# **Case** para vaga de cientista de dados

# Olá!

## Eu sou Gustavo Cunha

Cientista de dados na A3Data

Professor de ciência de dados no Le Wagon

AWS Certified Machine Learning Specialist

Ex-Oficial das Forças Armadas

https://www.linkedin.com/in/ds-gustavo-cunha/

# Descrição geral do case

- **Problema:** gerar insights com base nos dados disponibilizados.
- **Dados:** amostra de mais de 1,5 milhão de compras feitas na Hotmart em 2016.
- **Prazo e escopo:** 72h e escopo não limitado às perguntas iniciais.
- **Entregável:** apresentação com storytelling da solução e código no Github.

- **Perguntas iniciais:**
  - A Hotmart depende dos maiores produtores da plataforma?
  - Existe algum padrão ou tendência relevante nos dados?
  - É possível segmentar os usuários com base em suas características?
  - Quais características mais impactam no sucesso de um produto?
  - É possível estimar quanto de faturamento a Hotmart irá fazer nos próximos três meses a partir do último mês mostrado no dataset?

# SOLUÇÃO DO CASE

# Índice



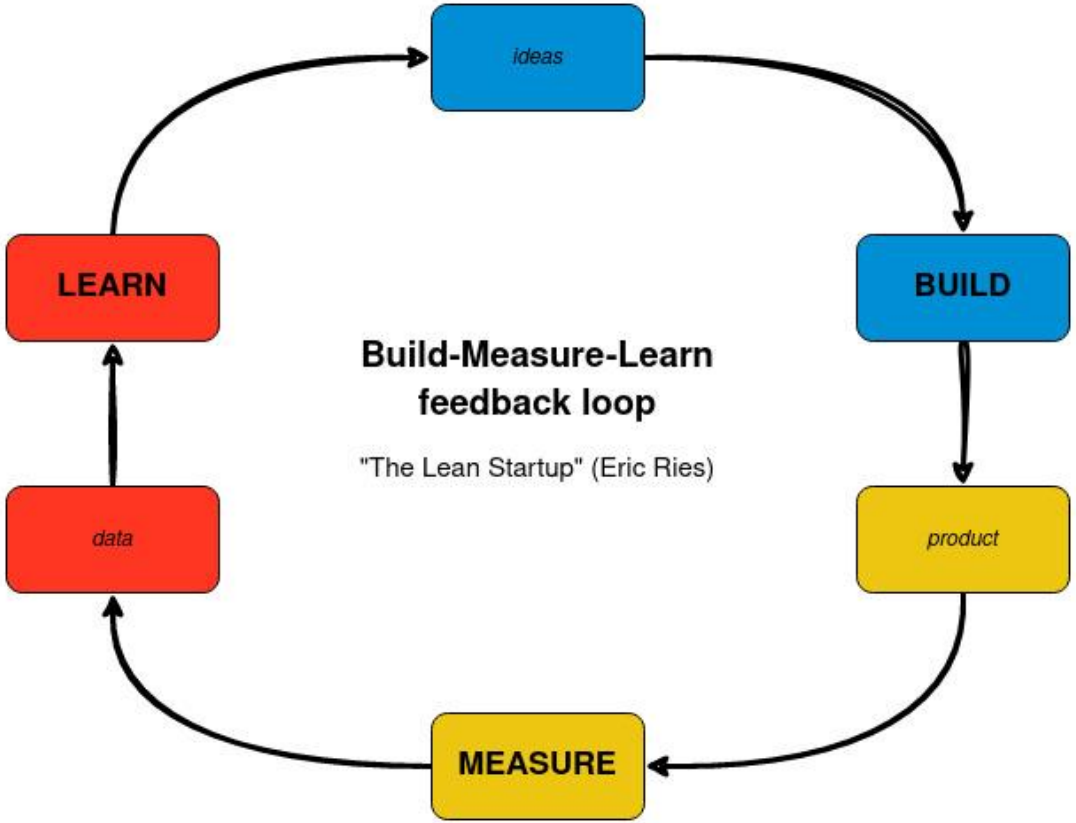
# 1. Situação

Contexto e problema de negócio

**SITUAÇÃO** TAREFA AÇÃO RESULTADO

# Design do problema

"Build-measure-learn" feedback loop


ideas
BUILD
product
MEASURE
data
LEARN
Build-Measure-Learn
feedback loop
"The Lean Startup" (Eric Ries)

# Contexto de negócio

**Qual a empresa?**

Hotmart

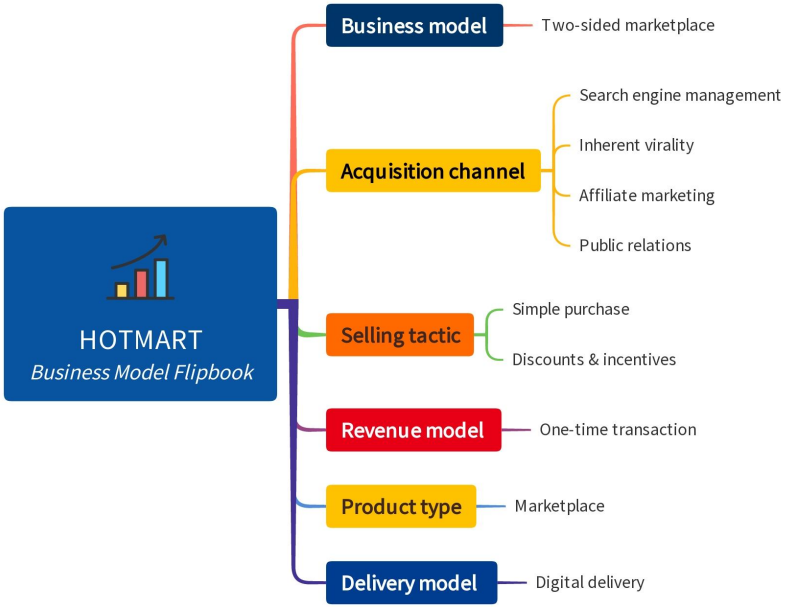

**Qual o modelo de negócio?**

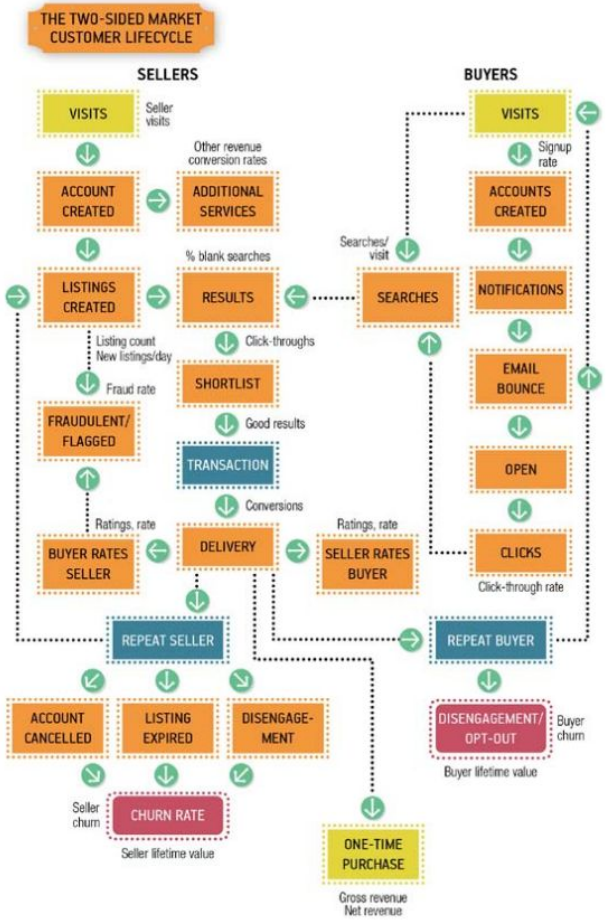
**Two-sided marketplace**. É uma plataforma de compra, venda e divulgação de produtos digitais em que a Hotmart conecta os criadores/divulgadores de produtos aos seus clientes.

**Qual o estágio do negócio?**

**Viralidade** (Lean Analytics) ou “**early majority**” (Innovation Adoption Curve). Encontrou uma dor no mercado e validou um produto que resolve a dor; agora é o momento de **aumentar a base de clientes**.

# Contexto de negócio*




*"Lean Analytics: Use Data to Build a Better Startup Faster" de Alistair Croll & Benjamin Yoskovitz

# Problema de negócio

**Qual problema de negócio a empresa está enfrentando?**

A empresa deseja obter **insights** com base nos dados dos clientes para revelar novas **oportunidades de produtos** (especialmente em termos de sucesso do produto, segmentação de clientes e estimativa de receita) de modo a dar **suporte** à **estratégia** de **aumentar** a **base de clientes** (escalar).

**Qual solução esse projeto deve entregar?**

Uma apresentação (ppt) com os **insights** obtidos por meio dos dados disponíveis e, possivelmente, respostas às perguntas de negócio.

# Escopo da solução e principais hipóteses assumidas

- **Business:**
  - O **foco** da empresa no momento é **escalar** a **base de clientes**.
  - Para **escalar**, é **mais importante** produtos que são **vendidos a mais clientes** e **comprados mais vezes** do que produtos que geram maior receita.

- **Dados:**
  - Quando o percentual de comissão do afiliado é nulo (ausente), considerou-se como sendo zero (0).
  - O valor de purchase_value está representado na escala z-score.

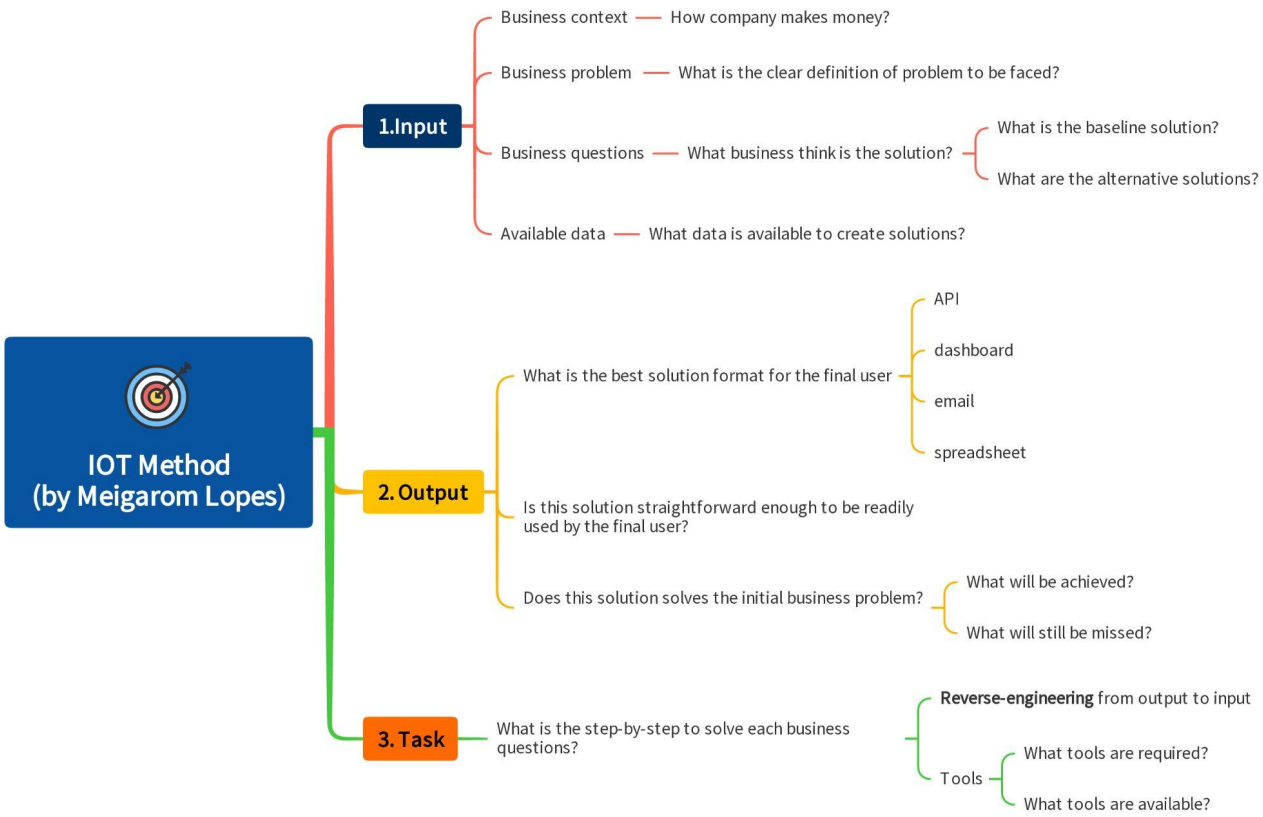
# 2. Tarefa

Roadmap do problema à solução

SITUAÇÃO | TAREFA | AÇÃO | RESULTADO

# Design da solução

Método IOT para criar o roadmap da solução


IOT Method
(by Meigarom Lopes)
1.Input
Business context — How company makes money?
Business problem — What is the clear definition of problem to be faced?
Business questions — What business think is the solution?
What is the baseline solution?
What are the alternative solutions?
Available data — What data is available to create solutions?
2. Output
What is the best solution format for the final user
API
dashboard
email
spreadsheet
Is this solution straightforward enough to be readily used by the final user?
Does this solution solves the initial business problem?
What will be achieved?
What will still be missed?
3. Task
What is the step-by-step to solve each business questions?
Reverse-engineering from output to input
Tools
What tools are required?
What tools are available?

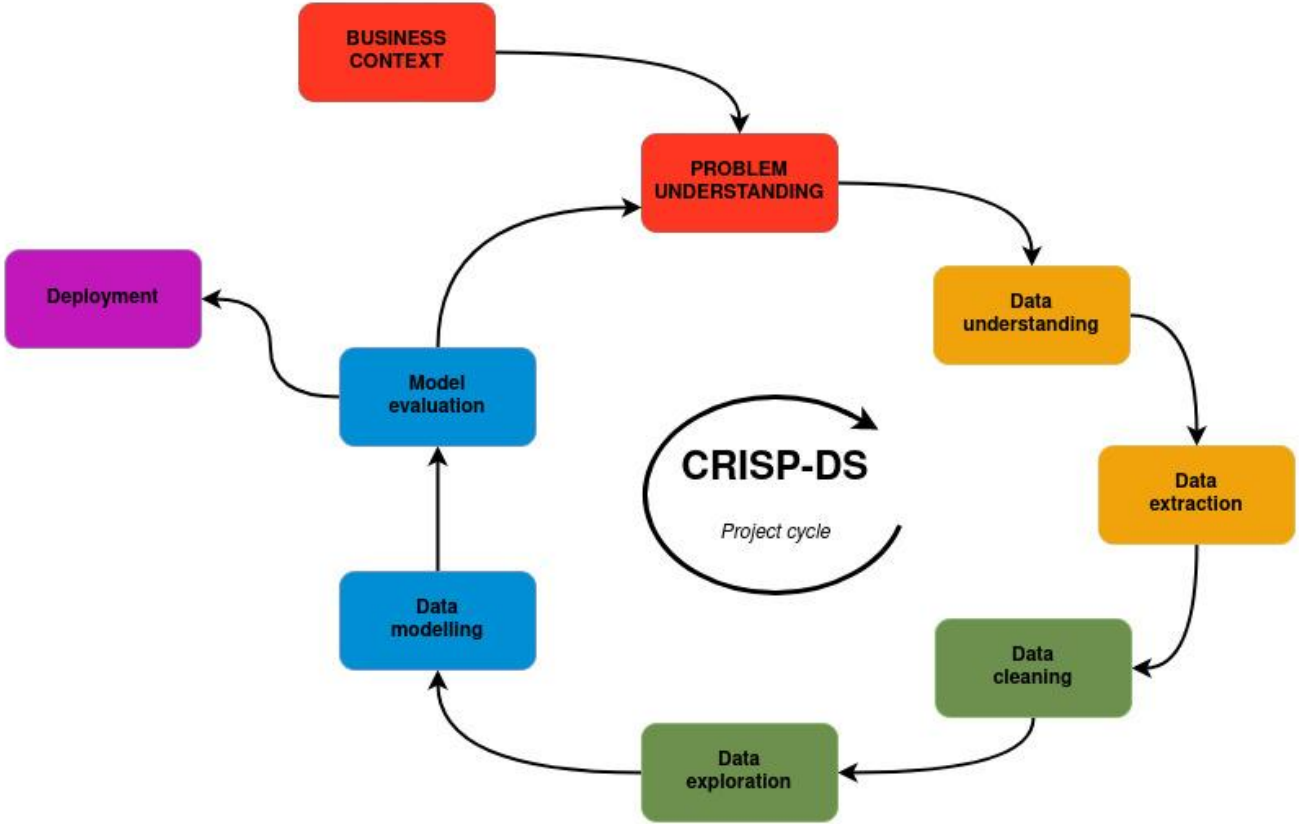
# 3. Ação

## Execução do roadmap de tarefas

SITUAÇÃO | TAREFA | AÇÃO | RESULTADO

# Produto de dados

CRISP-DS framework


BUSINESS CONTEXT
PROBLEM UNDERSTANDING
Data understanding
Data extraction
Data cleaning
Data exploration
Data modelling
Model evaluation
Deployment
CRISP-DS
Project cycle

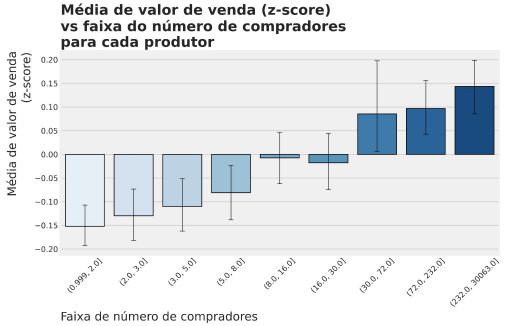
# PERGUNTA DE NEGÓCIO

A Hotmart **depende** dos **maiores produtores** da plataforma?

Os **produtores** que **mais vendem** são responsáveis pela **maior parte** do **faturamento** da Hotmart?

Para os **produtores**, explorar a relação entre:

- número de **clientes**;
- número de **produtos**;
- **valor** de **venda**



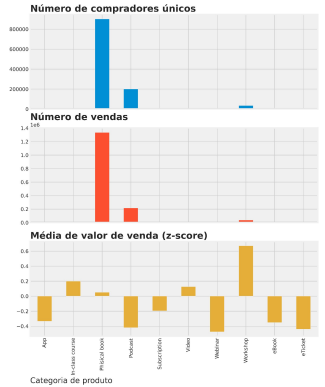
# PERGUNTA DE NEGÓCIO

Existe algum **padrão** ou tendência relevante nos **dados**?

Quais **características** mais impactam no **sucesso** de um **produto**? O que faz um **produto vender mais**?

Para os **produtos**, explorar a relação entre:

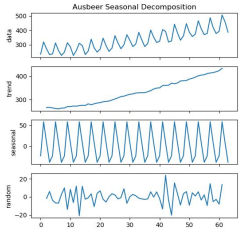
- número de **compradores** únicos;
- número de **vendas**;
- **valor** de **venda**



# PERGUNTA DE NEGÓCIO

É possível **estimar** quanto de **faturamento** a Hotmart irá fazer nos **próximos três meses** a partir do último mês mostrado no dataset?

*Seasonal decomposition*




$$SARIMA\underbrace{(p,d,q)}_{non-seasonal}\underbrace{(P,D,Q)_m}_{seasonal}$$

PROPHET

# PERGUNTA DE NEGÓCIO

É possível **segmentar**
os **usuários** com base
em suas características?

**R**ecency

**F**requency

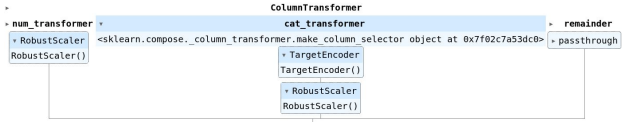
**M**onetary

# PERGUNTA DE NEGÓCIO

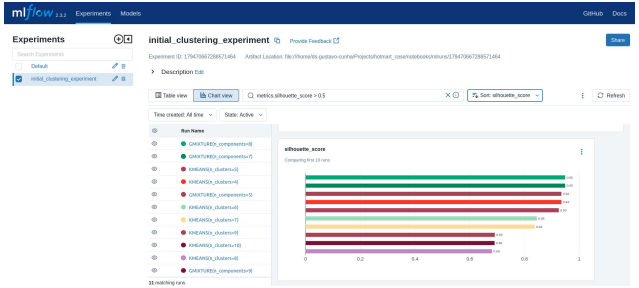
Existe algum **padrão** ou tendência relevante nos **dados**?

## Clusterização de produtos

- SKlearn pipelines



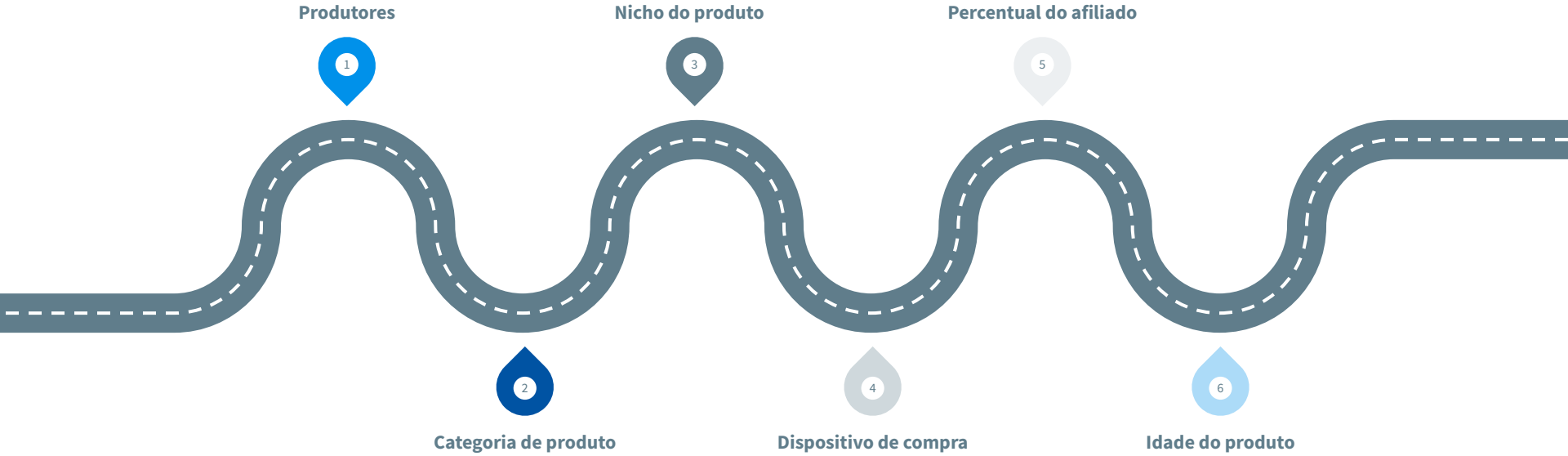

- MLFlow



# 4. Resultado

Principais insights e resultados

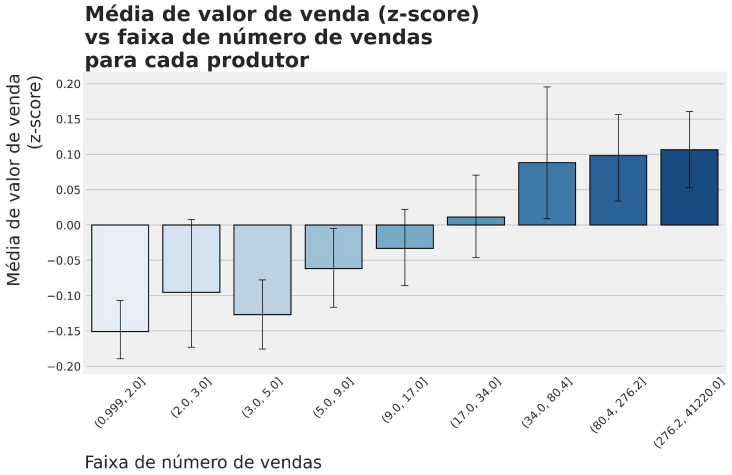
# Insights da análise exploratória de produto

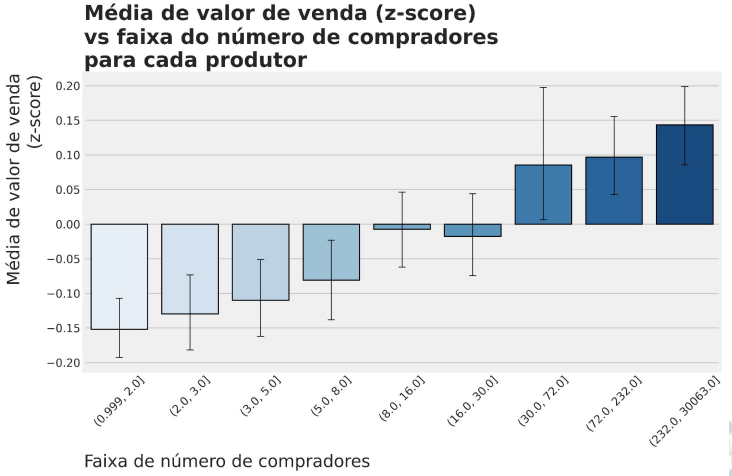


# PRODUTORES

- Produtores com mais produtos vendidos e com maior número de compradores únicos tendem a ter uma maior média de valor de vendas.

- **Tanto os produtores que produzem mais conteúdo quanto os que vendem para mais clientes têm um valor médio de venda maior.** *Provavelmente são uma peça chave para o negócio em termos de escalabilidade e de receita (trazem clientes e trazem receita).*

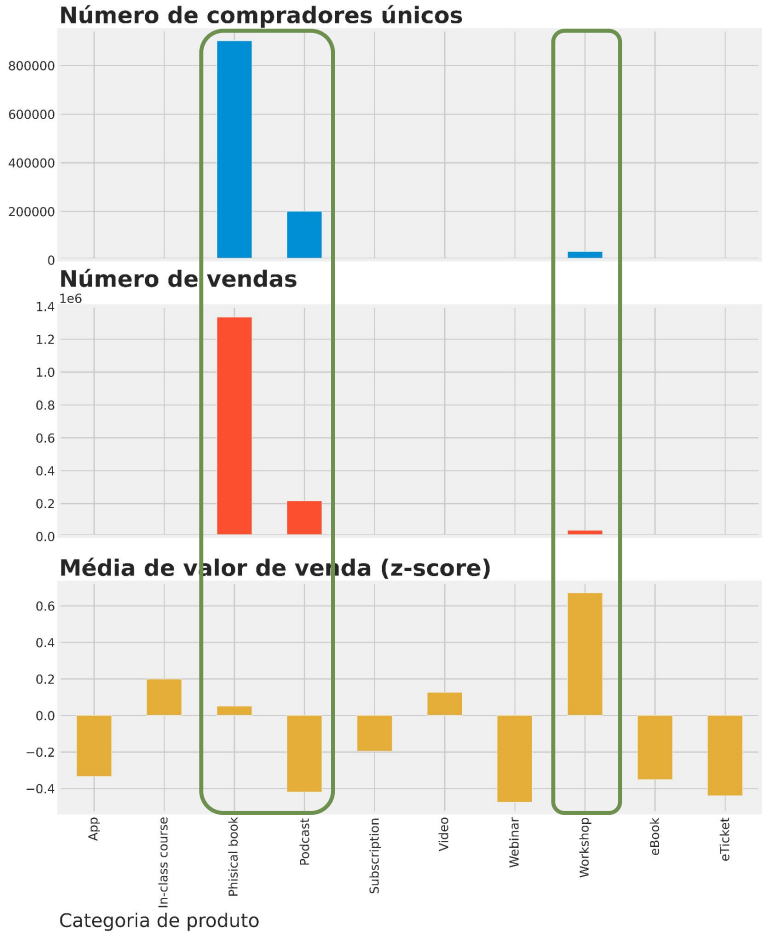


# CATEGORIA DO PRODUTO

- Livros físicos estão correlacionados a um valor de venda médio porém a alto número de vendas e de compradores.
- Podcasts estão correlacionados a um baixo valor de venda porém a maiores números de vendas e de compradores.
- Workshops estão correlacionados a um alto valor de venda porém a um baixo número de vendas e de compradores.

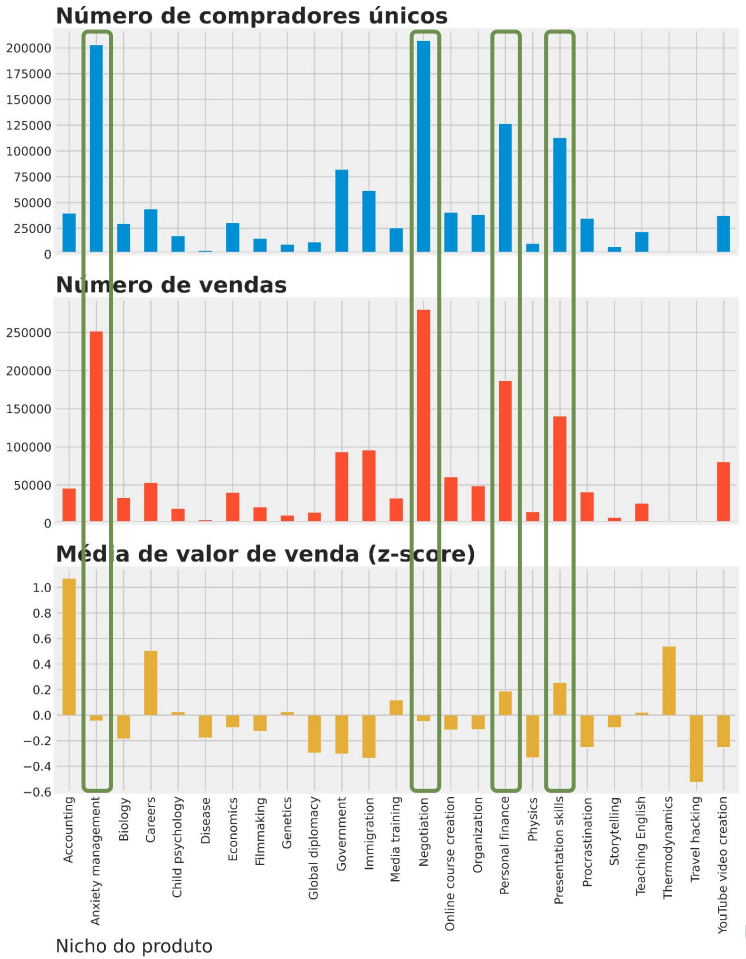
- **Mesmo tendo menor valor de venda, as categorias livros físicos e podcasts podem ser “melhores” que workshops para escalar.**



# NICHO DO PRODUTO

- Gerenciamento de ansiedade e negociação estão correlacionados a um baixo valor de venda porém a alto número de vendas e de compradores.
- Finanças pessoais e habilidades de apresentação estão correlacionados a um maior valor de receita, de número de vendas e de compradores.

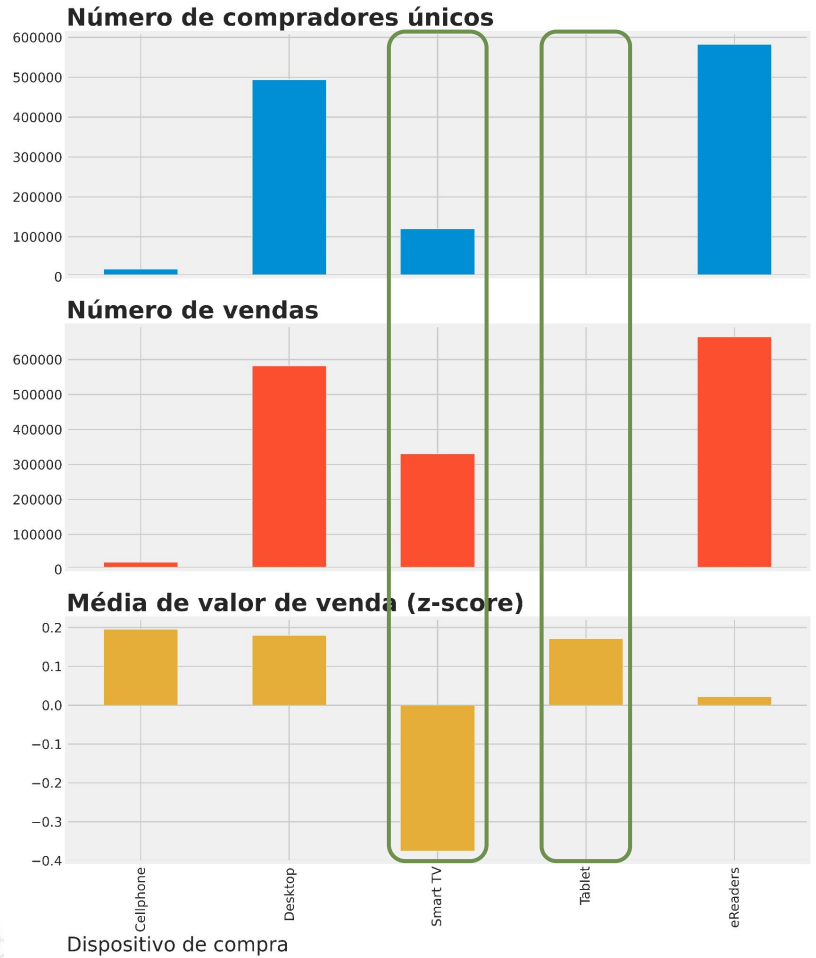
- **Mesmo não tendo valor de venda alto, os nichos de gerenciamento de ansiedade e negociação podem ser os “melhores” para escalar.**



# DISPOSITIVO DE COMPRA

- Smart TV está correlacionado a um baixo valor de venda porém a alto número de vendas e de compradores.
- Tablet está correlacionado a um alto valor de vendas porém a um número baixo de vendas e de compradores.

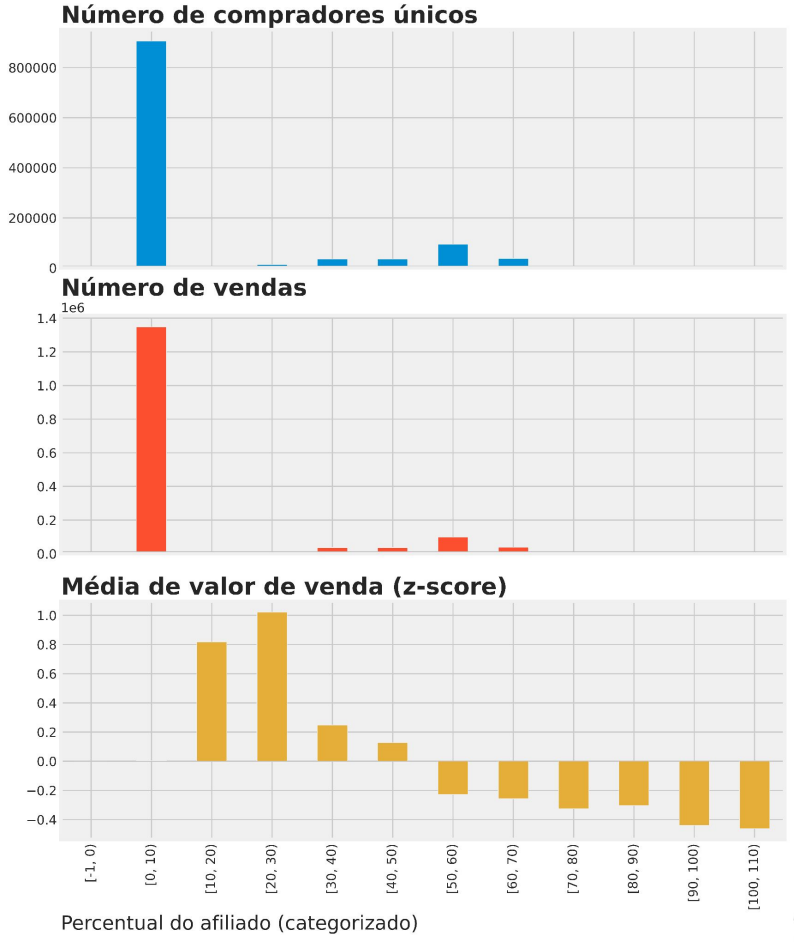
- **Mesmo tendo menor valor de venda, os dispositivos Smart TV podem ser “melhores” que os Tablets para escalar.**



# PERCENTUAL DE COMISSÃO DO AFILIADO

- Percentuais de 0 a 10% estão correlacionados a um valor de venda médio porém a alto número de vendas e de compradores.
- Percentuais de 10 a 50% estão correlacionados a um maior valor de venda porém a baixos números de vendas e de compradores.
- Percentuais maiores que 50% estão correlacionados a um baixo valor de venda, número de vendas e de compradores.

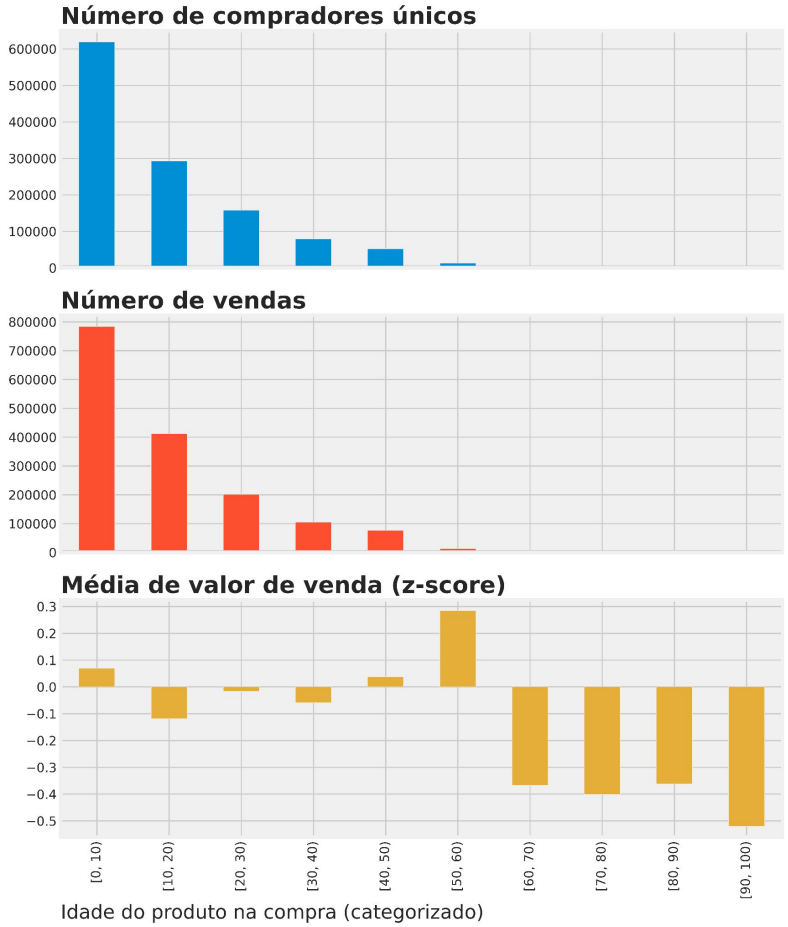
- **Produtos com percentual de comissão do afiliado de até 10% podem ser os “melhores” para escalar.**



# IDADE DO PRODUTO

- Produtos mais recentes estão correlacionados a maiores números número de vendas e de compradores.
- Produtos mais velhos estão correlacionados a baixos valores de vendas, números número de vendas e de compradores.

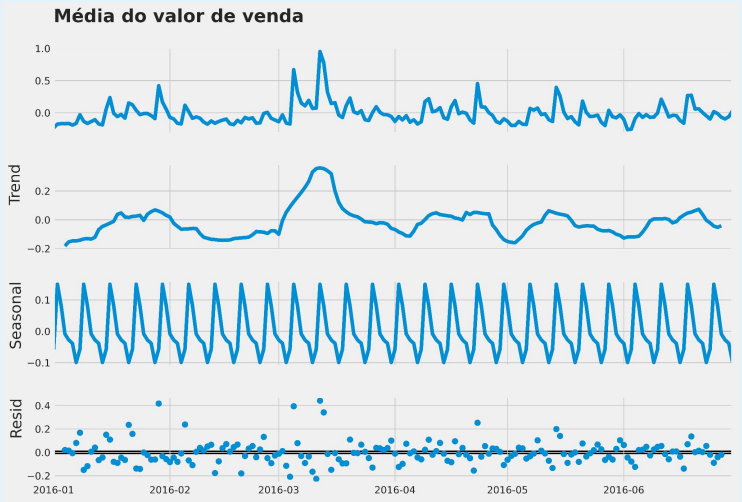
- **Produtos mais recentes podem ser os “melhores” para escalar.**



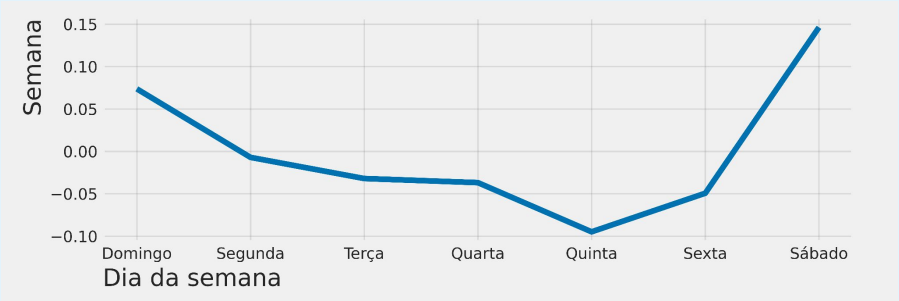
# Previsão de vendas

# EXPLORATÓRIA DA PREVISÃO DE VENDAS MÉDIA POR DIA

- Tendência não bem definida
- Sazonalidade semanal bem evidente




*Seasonal decomposition*



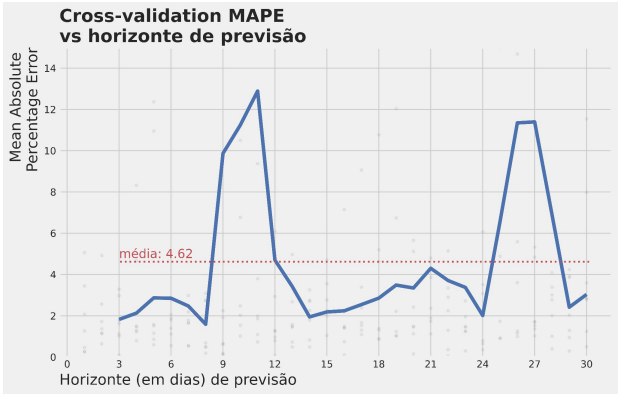

*Facebook Prophet*

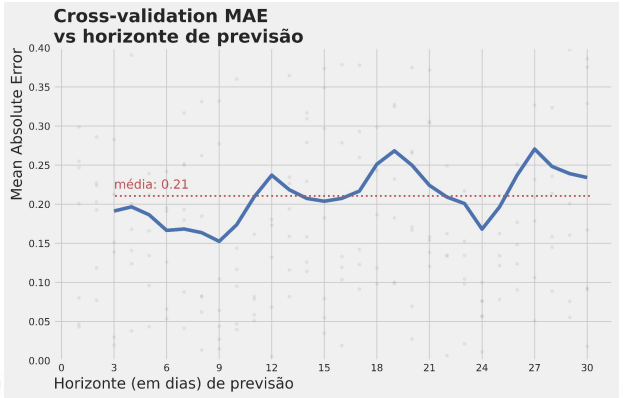
# PREVISÃO DE VENDAS MÉDIA POR DIA

## *Facebook Prophet*

- Média de vendas por dia (z-score)
- Horizonte de previsão: 30 dias

- Média CV MAPE: 4.62
- Média CV MAE: 0.21



# Recency-Frequency-Monetary

# Clusterização RFM (Recency-Frequency-Monetary)

Streamlit app

- Cluster **VERÃO**:
  - Número de compradores: **11.236** (1.02% do total)
  - Recência média: **13 dias** (desde última compra)
  - Frequência média: **6.7 compras**
  - Média de valor de compra: **0.5** (**z-score**)
  - **Business: muito pouco compradores, R alto, F alto e M alto**

- Cluster **PRIMAVERA**:
  - Número de compradores: **182.199** (16.55% do total)
  - Recência média: **25 dias** (desde última compra)
  - Frequência média: **2.5 compras**
  - Média de valor de compra: **0.6** (**z-score**)
  - **Business: poucos compradores, R médio, F médio e M alto**

- Cluster **OUTONO**:
  - Número de compradores: **511.019** (46.43% do total)
  - Recência média: **69 dias** (desde última compra)
  - Frequência média: **1.3 compras**
  - Média de valor de compra: **0.2** (**z-score**)
  - **Business: maioria dos compradores, R baixo, F baixo e M médio**

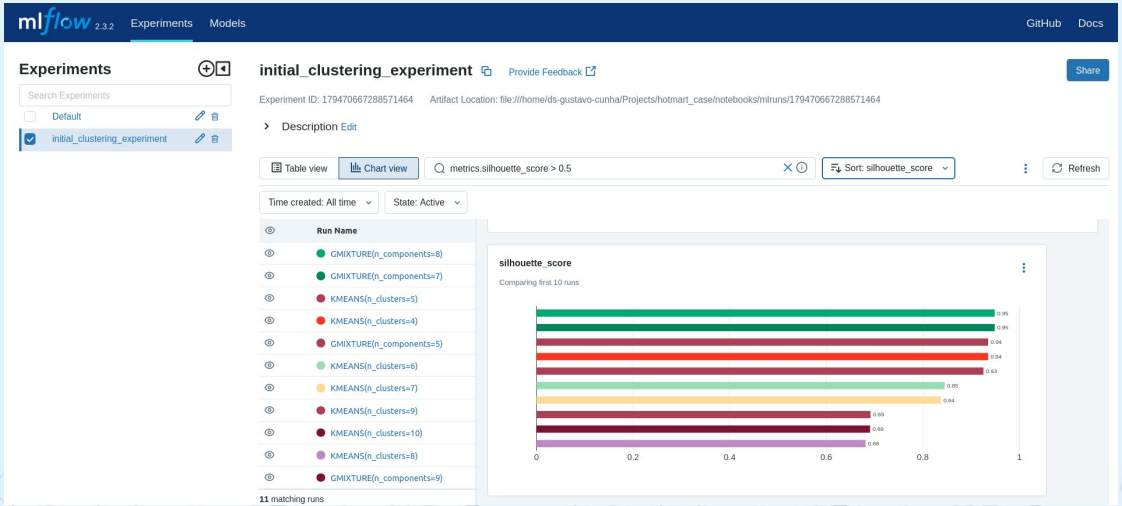
- Cluster **INVERNO**
  - Número de compradores: **396.195** (36.00% do total)
  - Recência média: **120 dias** (desde última compra)
  - Frequência média: **1.1 compras**
  - Média de valor de compra: **-0.4** (**z-score**)
  - **Business: muitos compradores, R baixo, F baixo e M baixo**

# Segmentação de produto

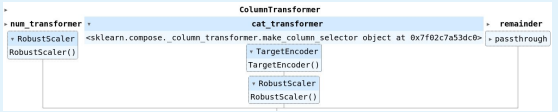
# Clusterização de produtos

- Métrica de clusterização: **silhouette score**
- Escolha do algoritmo: trade-off entre silhoutte (score e shape) e tamanho dos clusters
- Algorítmo escolhido:
  - **KMeans**(n_clusters=8)
- Número final de **clusters**: **4**
  - Clusters menores foram agrupados

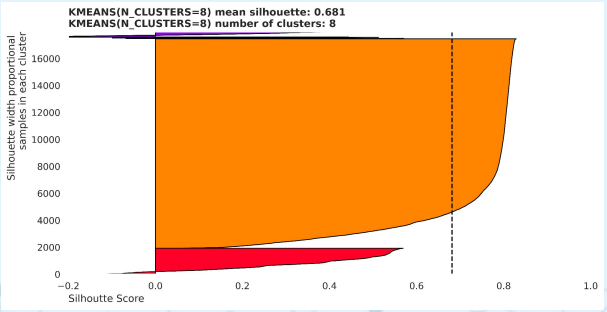
*MLFlow*




*SKLearn pipelines*




*Clusters silhouette*



## Cluster **SOL**:

- número de **produtos** no cluster: **76**
- média de **compradores** únicos: **5039**
- média de número de **compras**: **6980**
- média de **valor** de **compra** (z-scored): **-0.13**
- média de número de afiliados: 122
- nicho mais frequente: Anxiety management
- média do percentual de comissão do afiliado: 10.08
- **Business:** ***"poucos produtos; não traz receita mas traz muitas compras e compradores"***

## Cluster **MERCÚRIO**:

- número de **produtos** no cluster: **361**
- média de **compradores** únicos: **947**
- média de número de **compras**: **1104**
- média de **valor** de **compra** (z-scored): **0.09**
- média de número de afiliados: 31
- nicho mais frequente: Anxiety management
- média do percentual de comissão do afiliado: 13.23
- **Business:** ***"poucos produtos; equilibrado em receita, compras e compradores"***

## Cluster **VÊNUS**:

- número de **produtos** no cluster: **15534**
- média de **compradores** únicos: **33**
- média de número de **compras**: **38**
- média de **valor** de **compra** (z-scored): **0.11**
- média de número de afiliados: 1
- nicho mais frequente: Negotiation
- média do percentual de comissão do afiliado: 1.17
- **Business:** ***"maioria dos produtos; poucas compras e compradores, mas receita um pouco maior"***

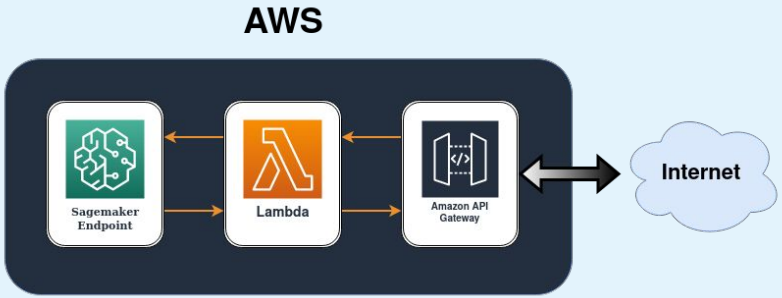
## Cluster **TERRA**:

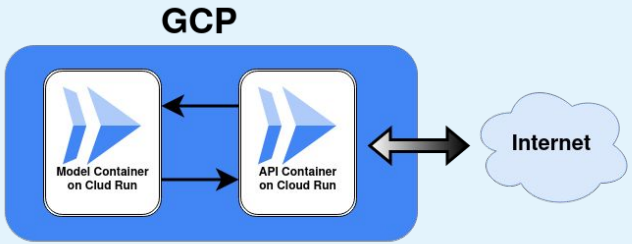
- número de **produtos** no cluster: **1911**
- média de **compradores** únicos: **37**
- média de número de **compras**: **37**
- média de **valor** de **compra** (z-scored): **0.17**
- média de número de afiliados: 5
- nicho mais frequEnte: Negotiation
- média do percentual de comissão do afiliado: 35.32
- **Business:** ***"número considerável de produtos; poucas compras e compradores, mas receita maior de todas"***

# Arquiteturas de deployment

# PREDIÇÃO ONLINE

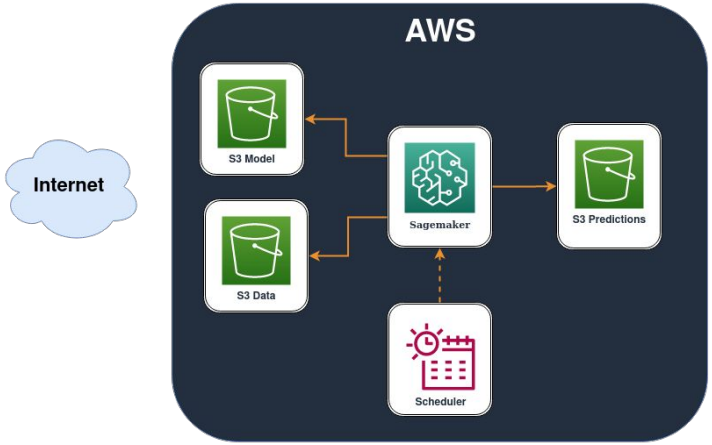
*Ex.: previsão de vendas diárias*




# PREDIÇÃO EM BATCH

*Ex.: clusterização mensal de produtos*



# Próximos passos

- **Iterar** mais um ciclo do **CRISP-DM** nos pontos onde há **maior entrega** de **valor** ao **business.**
- **Validar** entendimento dos **insights** com time de negócio.
- *Aprofundar* exploratória e modelagem de *série temporal* para estimar faturamento.
- *Validar* o valor para o negócio e a utilização da *segmentação RFM*.
- *Aprofundar* a *clusterização* de produto para melhorar a distinção entre clusters bem como definir o consumo dessa clusterização.
- *Definir arquitetura* de *deployment* para modelo de previsão de faturamento e para modelo de clusterização de produto.

# Código no Github

# Muito Obrigado!

## Alguma pergunta?

Linkedin: https://www.linkedin.com/in/ds-gustavo-cunha/

Portfolio: https://ds-gustavo-cunha.github.io/projects-portfolio/

Email: gcunhaj@gmail.com

# Credits

- Template de apresentação: SlidesCarnival
- Fotos: Unsplash