

# Big Data Analytics com R e Microsoft Azure Machine Learning



Data Science Academy



# Regressão



Data Science Academy



# Processo de Data Science (Big Data Analytics)

- Compreender o Problema a ser Resolvido
- Coletar os Dados
- Limpar, Compreender e Preparar os Dados
- Selecionar e Transformar as Variáveis
- Construir, Testar, Avaliar e Otimizar o Modelo
- Contar a História dos Dados



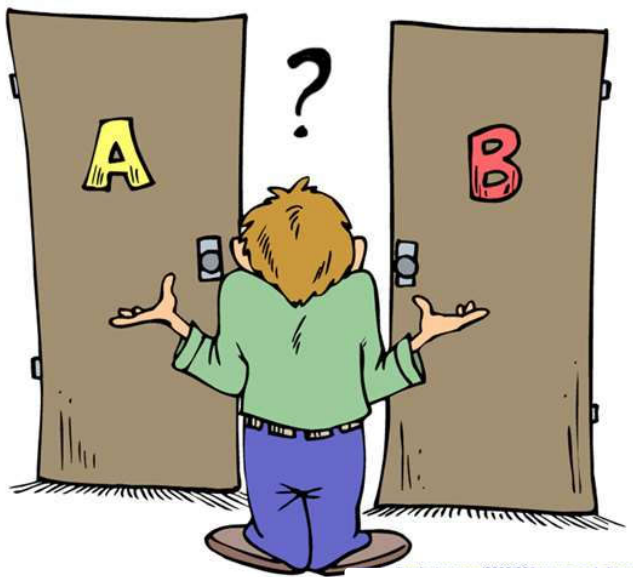
Data Science Academy

# Feature Selection



Data Science Academy

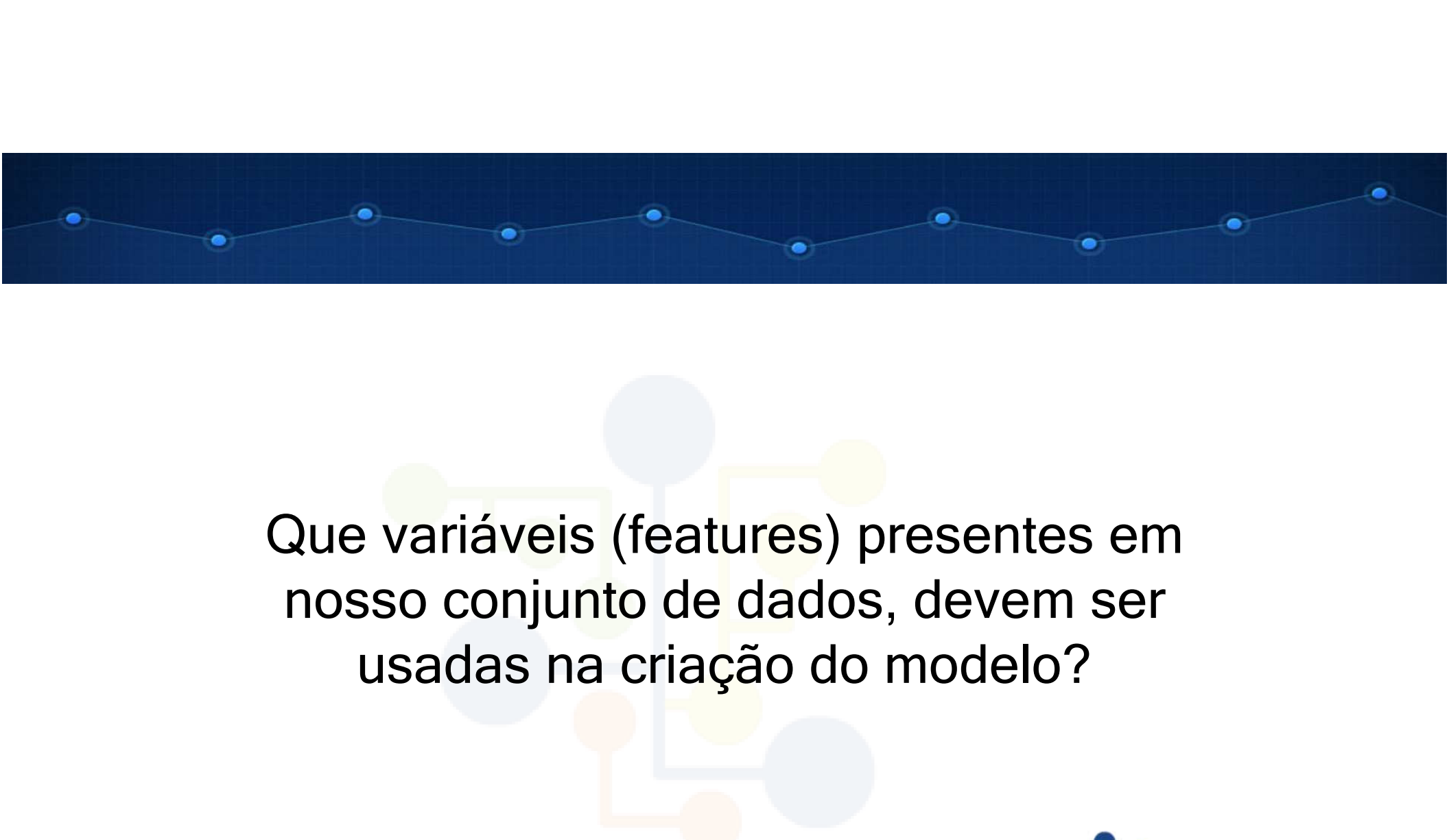
# Feature Selection



- Simplificação do Modelo, para facilitar sua interpretação
- Reduzir o tempo de treinamento do modelo
- Melhorar a generalização do modelo, evitando overfitting



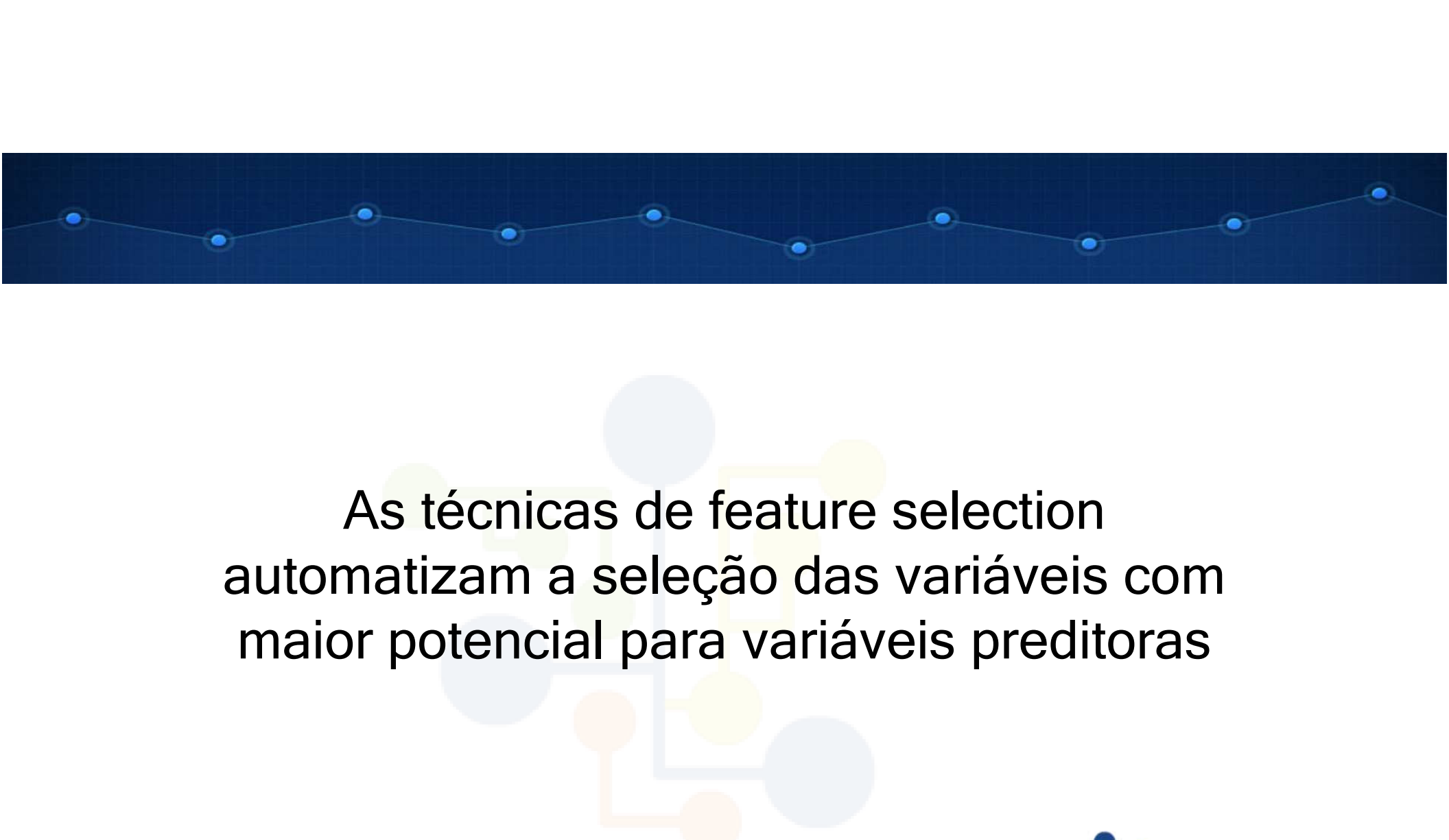
Data Science Academy



Que variáveis (features) presentes em  
nosso conjunto de dados, devem ser  
usadas na criação do modelo?




Data Science Academy



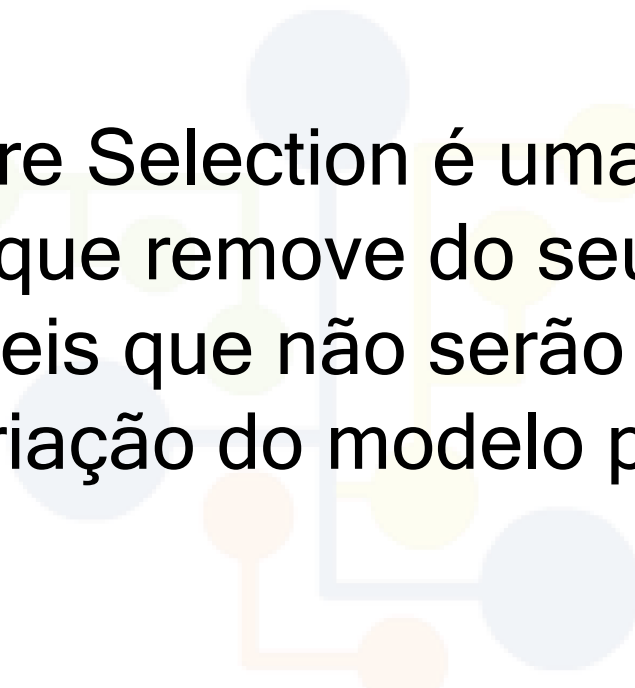
As técnicas de feature selection automatizam a seleção das variáveis com maior potencial para variáveis preditoras



Data Science Academy

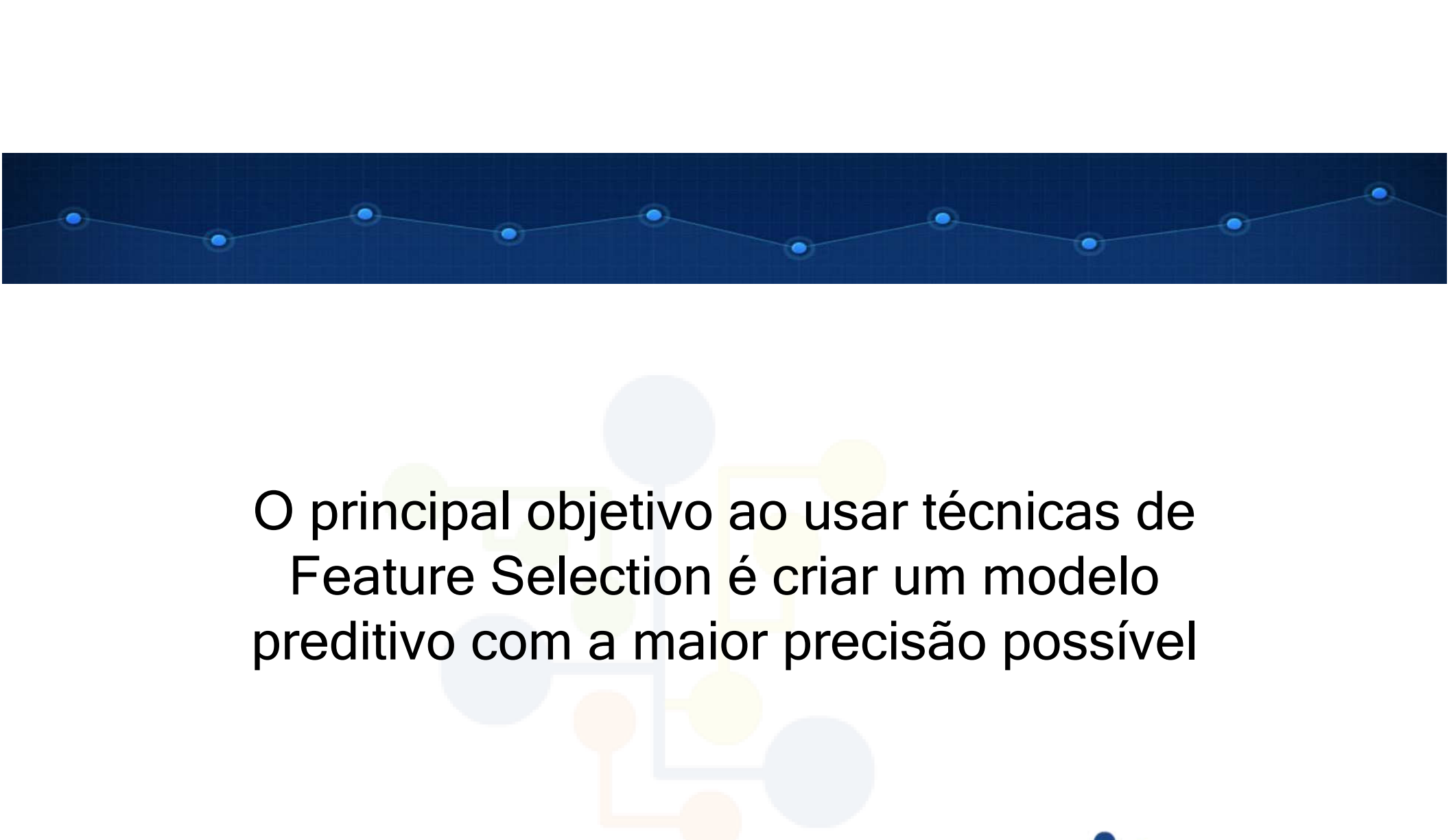


Feature Selection é uma espécie de filtro, que remove do seu dataset as variáveis que não serão úteis para a criação do modelo preditivo



Data Science Academy





O principal objetivo ao usar técnicas de Feature Selection é criar um modelo preditivo com a maior precisão possível



Data Science Academy



# Existem diversos métodos para Feature Selection

Teste do Qui-quadrado

Coeficientes de Correlação

Algoritmos de Eliminação Recursiva

Algoritmos de Regularização (LASSO, Elastic Net, Ridge Regression)



Data Science Academy



Feature Selection



Redução de Dimensionalidade



Data Science Academy



# Redução de Dimensionalidade

Principal Component Analysis (PCA)

Singular Value Decomposition (SVD)



Data Science Academy



## Antes de aplicar Feature Selection, diversas perguntas devem ser respondidas

- Suas variáveis são mensuráveis?
- Você encontrou interdependência entre as variáveis?
- Você tem conhecimento sobre a área de negócio que gerou os dados?
- Sabe identificar as variáveis mais relevantes dentro do seu conjunto de dados?
- A análise exploratória dos dados encontrou "sujeira" nos seus dados?



Data Science Academy



Basicamente, calculamos o nível de significância de cada variável e eliminamos aquelas com significância mais baixa



Data Science Academy



Veremos duas formas de fazer isso

Usando o módulo **Filter Based Feature Selection** do Azure ML

Criando um **modelo randomForest** para calcular a significância de cada variável, usando R



Data Science Academy


# Storytelling



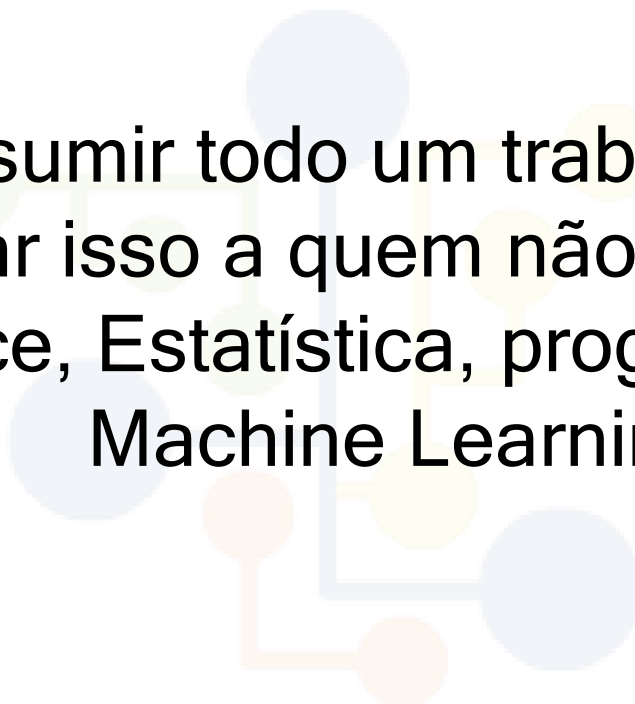
Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)





Como resumir todo um trabalho de análise e explicar isso a quem não conhece Data Science, Estatística, programação ou Machine Learning?



Data Science Academy



Não utilize linguagem técnica em apresentações executivas



Data Science Academy

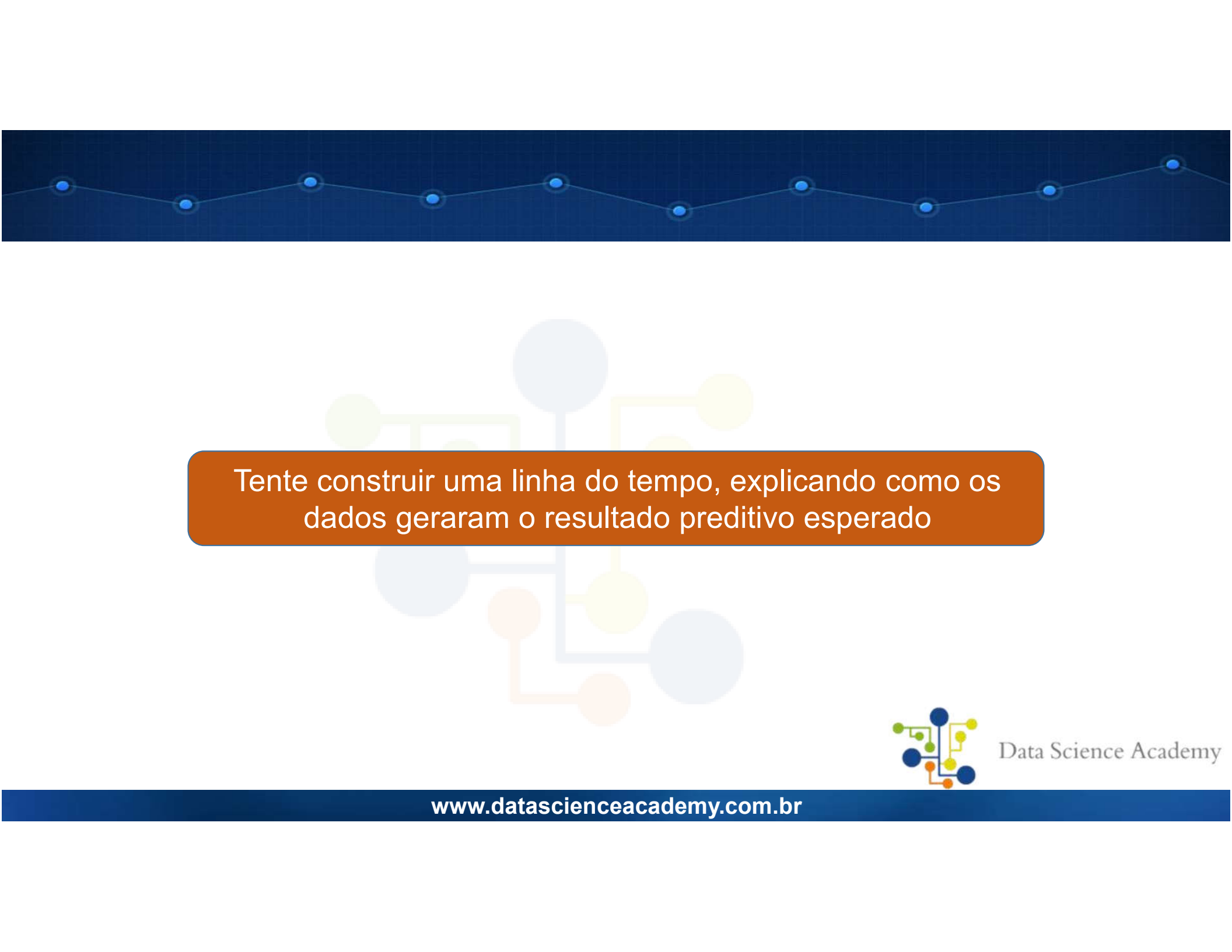
[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Use apenas um ou dois gráficos para explicar seus resultados (use tabelas se necessário)



Data Science Academy



Tente construir uma linha do tempo, explicando como os dados geraram o resultado preditivo esperado



Data Science Academy



Não use termos como "variáveis" e sim  
"atributos" ou "características"



Data Science Academy




Documente todo seu trabalho, pois isso servirá de apoio  
para que você possa contar a história dos dados

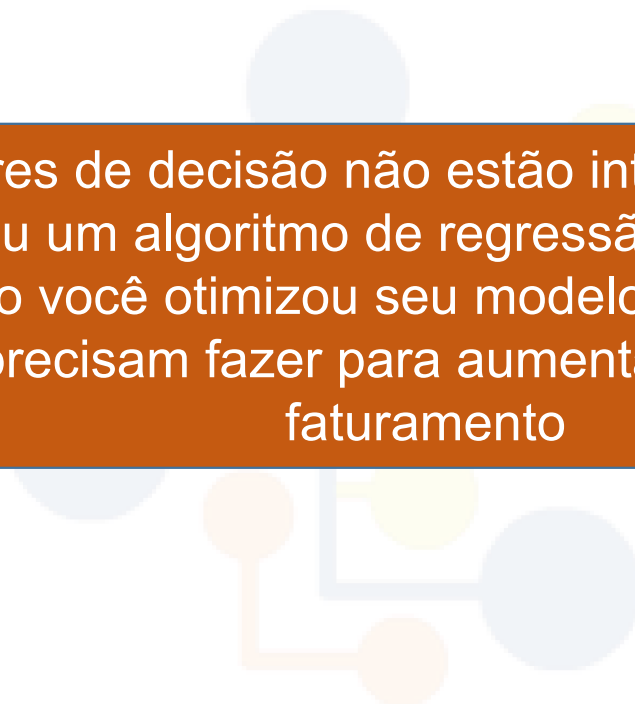


Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Os tomadores de decisão não estão interessados em saber se você usou um algoritmo de regressão ou de classificação e nem como você otimizou seu modelo. Eles querem saber o que precisam fazer para aumentar as vendas e o faturamento



Data Science Academy

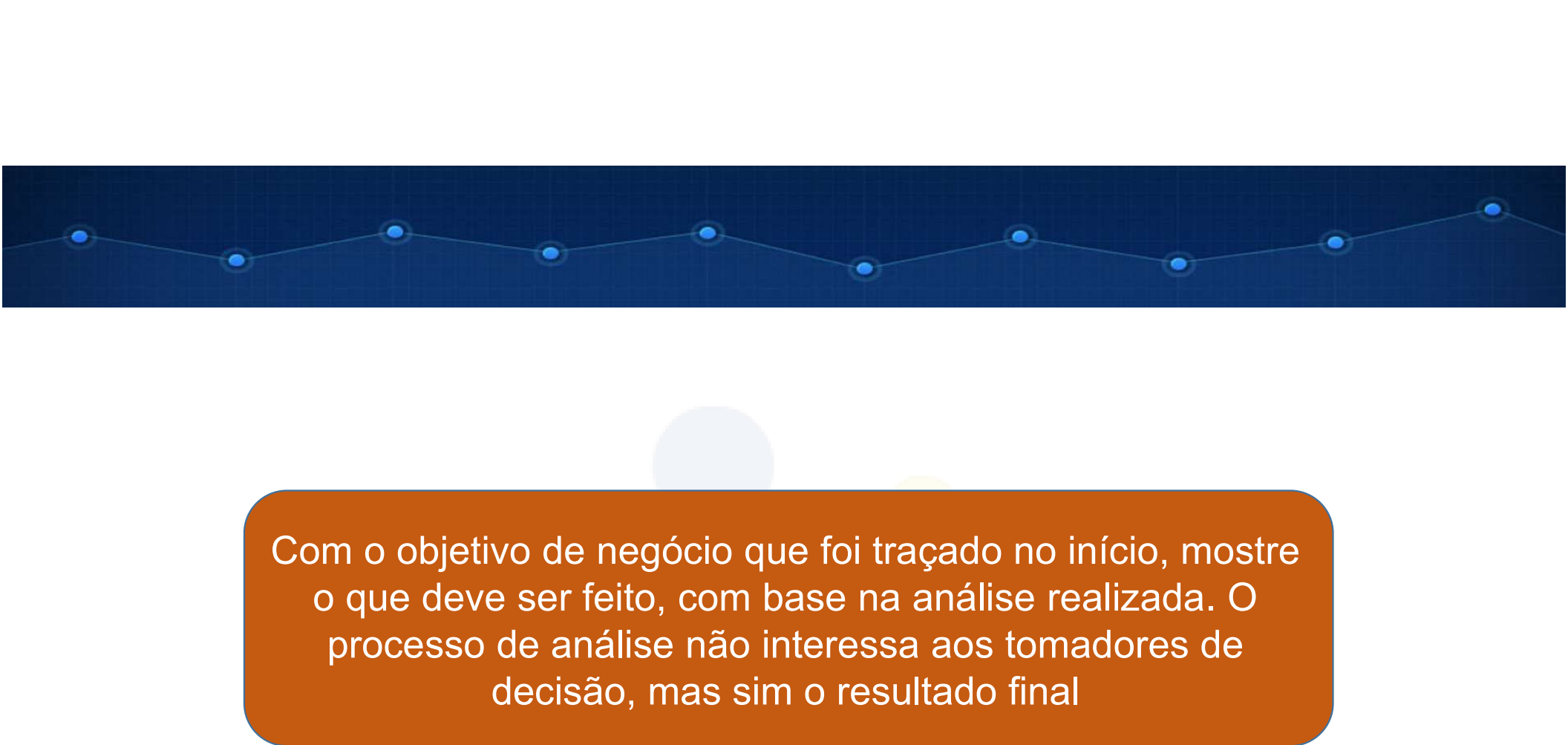


Responda o que é importante para eles e não o que é importante para você



Data Science Academy






Com o objetivo de negócio que foi traçado no início, mostre o que deve ser feito, com base na análise realizada. O processo de análise não interessa aos tomadores de decisão, mas sim o resultado final



Data Science Academy



A fase de Análise Exploratória dos Dados pode apresentar insights preciosos. Use isso ao seu favor e guarde todos os resultados intermediários do seu processo de análise!



Data Science Academy



Storytelling é uma arte e requer prática para ser exercida  
com maestria



Data Science Academy