# 10 passos para um projeto de Data Science:

1. Descrição dos dados
2. Feature Engineering
3. Filtragem de dados e registros (de acordo com o negócio)
4. Exploração dos dados (2 Objetivos: Entender o negócio e gerar insights, e encontrar a relevância das variáveis para a construção do modelo)
5. Preparação dos dados: Enconding das variáveis categóricas e numéricas e transformação das variáveis
6. Filtragem de variáveis que são mais relevantes pelo modelo (Boruta e EDA)
7. Implementação de algoritmos de ML (Baseline, Lineares e não-lineares)
8. Hiperparâmetro Fine Tuning.

# Entendimento do negócio

## Contexto:

A Rossmann opera mais de 3.000 drogarias em 7 países europeus. Atualmente, os gerentes de loja da Rossmann têm a tarefa de prever suas vendas diárias com até seis semanas de antecedência. As vendas da loja são influenciadas por muitos fatores, incluindo promoções, competição, feriados escolares e estaduais, sazonalidade e localidade. Com milhares de gerentes individuais prevendo vendas com base em suas circunstâncias únicas, a precisão dos resultados pode ser bastante variada.

## Questão de negócio

Qual é o valor das vendas de cada loja nas próximas 6 semanas?

## Qual a motivação?

A previsão de vendas foi requisitada pelo CFO em uma reunião mensal sobre os resultados das lojas.

## Qual a causa raiz do problema?

Dificuldade em determinar o valor do investimento para as reformas de cada loja.

## Quem é o dono do problema?

Diretor financeiro (CFO) da Rossman.

## Qual é o formato da solução?

**Granularidade:** Previsão de vendas por dias e por loja os próximos 42 dias (6 Semanas)

**Tipo do problema:** Previsão de vendas (Regressão)

**Potenciais métodos:** Séries temporais e regressão com algumas modificações.

**Formato de entrega (3 itens):**

    - O valor total das vendas no final das 6 semanas (Uma coluna com o código ID da loja e outra coluna com o valor de vendas).

    - A entrega será pelo celular (app).

    - Checagem diária.

# Ciclo 01

## ENTENDIMENTO DO NEGÓCIO

Objetivo, cliente, formato de solução e frequência da solução.

## COLETA DE DADOS

Entender como vai ser a ingestão dos dados ao ser uma implementada uma aplicação e entender como obter os dados de treinamento e validação para a formulação do algoritmo.

## VISUALIZAÇÃO DOS DADOS

Visualizar os dados no primeiro momento para entender quais campos existem, podendo ser utilizado o excel para isso.

## DESCRIÇÃO DOS DADOS

Preciso entender a dimensão das dificuldades que poderá ser encontrada ao realizar o projeto (Formato dos dados, volume dos dados, volume de dados faltantes)

* Data Visualization
* Rename columns
* Data Describe
* Modifiel Type: Modificação da coluna ‘date’ para datetime
* Check NA: Analisando a representatividade dos dados faltantes,

## ATIVIDADES DESENVOLVIDAS NA BASE

* Renomeação das colunas
* Formatação do campo de data

## FEATURE ENGINEERING

Construção do mapa mental de hipóteses utilizando o Coggle. Devem ser descobertos os seguintes itens:

* Fenômeno:

    - Qual fenômeno estou modelando?

    Ex:

        - **\*\*Venda\*\***

* Agentes:

    - Quem são os agentes que atuam sobre o fenômeno de interesse?

    Ex:

        - Clientes

        - Lojas

        - Produtos

* Atributos dos agentes:

    - Descrição dos agentes

* Lista de Hipóteses:

    - Hipóteses para validar com os dados.

\* Duas formas de gerar insights:

    - Surpresa

    - Contraposição de crença

Hipóteses são apostas. Ex: Lojas de maior porte, deveriam vender mais... Lojas com mais sortimento, deveriam vender mais ...

* Criação de novas Features:
  + Year: Ano referente a data de venda
  + Month: Mês referente a data de venda
  + Day: Dia referente a data de venda
  + Week of year: Semana do ano referente a data de venda
  + Competition Since: Composição da data de competição
  + Competition since month: Composição dos meses que está competindo com outras lojas
  + Promo since: Composição de quanto tempo houve promoção nas lojas
  + Assortment: Mapeando as classes por extenso
  + State holliday: Mapeando as classes por extenso

## FILTRAGEM DE VARIÁVEIS:

Motivação: **RESTRIÇÕES DE NEGÓCIO. Tem que ser analisado no começo do projeto.**

Filtrar variáveis é diretamente relacionado às restrições do negócio, ou seja, as restrições são feitas de acordo com o negócio.

* Seleção de variáveis:
  + As variáveis mais relevantes do modelo. Pensar nas restrições que o time de negócio tem em obter os dados para o modelo em produção.

## EDA – Exploratory Data Analytics

Regra geral: Como as variáveis impactam o fenômeno? E qual a força desse impacto?

3 Objetivos de EDA:

* 1 – Ganhar experiência de negócio.
* 2 – Validar hipóteses de negócio (Gerar Insights: Pode ser por meio do choque de crenças ou pode ser por meio do fornecimento de uma informação nova causando surpresa).
* 3 – Perceber variáveis que são importantes para o modelo.

Análise univariada:

* Como é a distribuição dessa variável? (Ex: sns.countplot() ou distplot())
* Quais as medidas de tendência central e dispersão? (Ex: sns.boxplot())

Análise Bivariada:

* Como a variável impacta na resposta? (Ex: sns.barplot)
* Qual a linha de tendência? (Ex: sns,regplot)
* Correlação. (Ex: sns.heatmap(df.corr()))
* Validação de hipótese.

Análise Multivariada:

* Como as variáveis se relacionam?

## Preparação dos dados

O aprendizado da maioria dos algoritmos de ML é facilitado com dados numéricos, na mesma escala.

* Transformar dados categóricos em numéricos:
  + Conversão de features categóricas para numéricas (Encoding)- Existem vários tipos de encoding: One hot encoding (Funciona para níveis de estado), Label encoding (Não existe relação entre as categorias), Ordinal encoding (Existe uma relação ordinal entre as categorias), Target encoding (Leva em consideração a variável target para determinada categoria, funciona muito bem para muitos níveis de categorias), Frequency encoding e Embedding Encoding (Utilizada dentro de redes neurais profundas, geralmente quando se trata sobre NLP).
  + Transformação de natureza.
* Deixar os dados numéricos na mesma escala:
  + Normalização: Rescala o centro para 0 com desvio-padrão igual a 1 (Muito utilizado em dados com distribuição normal/gaussiana).
  + Rescaling: Rescala para o intervalo entre 0 e 1 (**Muito utilizado em distribuições não gaussianas)**.

- Normalização: (variável-média)/Desvio padrão (Pacote do scikit-learn)

- Rescaling: xi – média / (xmax – xmin) (O Min-max scaler é sensível aos outliers). Para variáveis com outliers pode-se utilizar o robust scale.

**Tipos de transformação:**

Transformação de grandeza: Trazer sua variável resposta o mais próximo possível de uma distribuição normal. A maioria dos algoritmos foram desenvolvidos em cima de algumas premissas, e uma dessas premissas é a distribuição normal da variável resposta.

Transformação de natureza: Trazer as variáveis mais próximas de sua real natureza. Alguns exemplos seria variáveis com natureza cíclicas. Ao utilizar a transformação de seno e cosseno na variável, aumenta a dimensionalidade acrescentando mais uma coluna, entretanto, o ganho de informação é maior do que essa por causa do aumento de dimensionalidade.

Exemplos:

1. Transformação logarítimica (Transformação de grandeza)
2. Box-Cox Trasformation (Transformação de grandeza)
3. Cube-root Transformation (Transformação de grandeza)
4. Square-root Transformation (Transformação de grandeza)
5. Sine and Cosine Transformation (Transformação de natureza)

## SELEÇÃO DE VARIÁVEIS

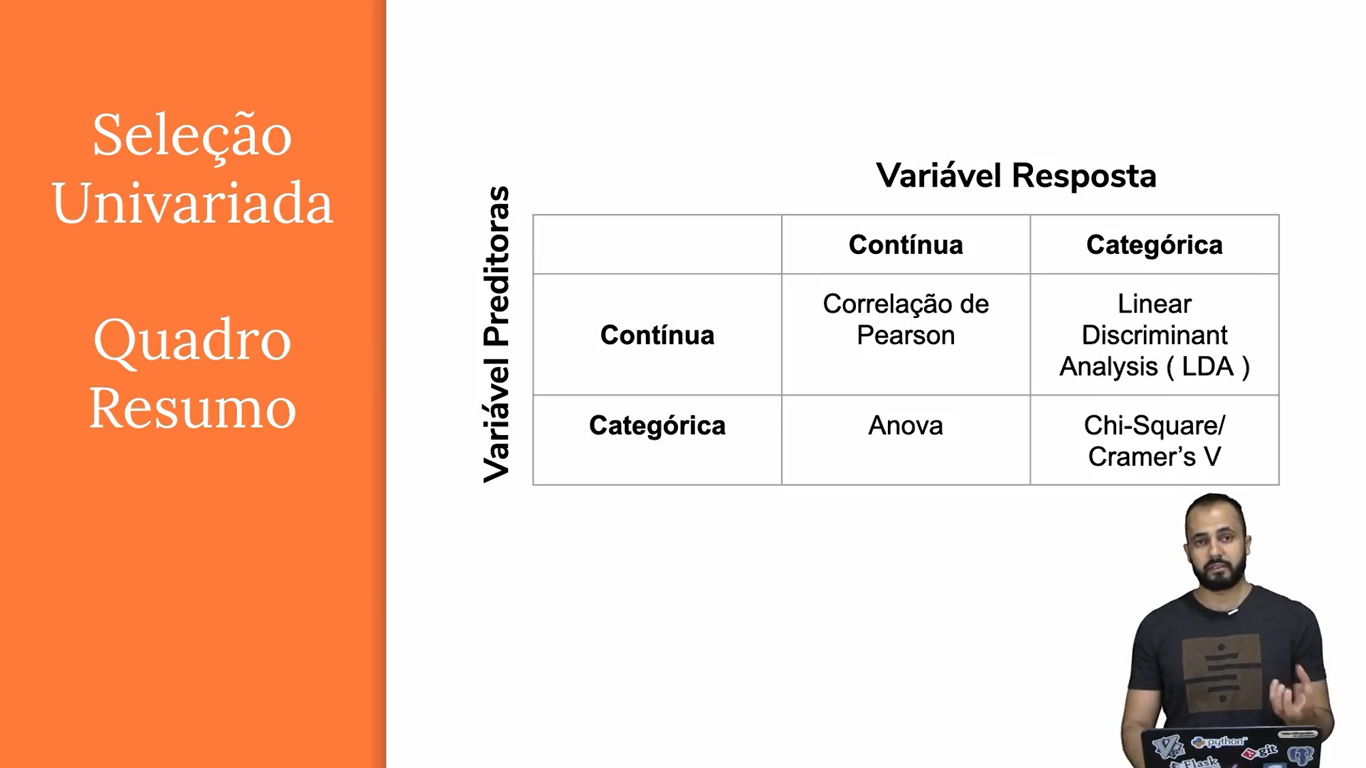
A motivação da seleção de variáveis vai de encontro com um princípio estatístico. Esse princípio é denominado de Navalha de Occam: “A explicação mais simples sobre um fenômeno observado, deveria prevalecer sobre explicações mais complexas.”

Dê sempre preferência para modelos mais simples....

Modelos mais simples aprendem melhor com um conjunto de dados mais simples. Ou seja, é necessário remover variáveis colineares por possuir o mesmo grau de informações (Explicam a mesma parte do fenômeno).

**Tipos de seleção de variáveis:**

1. **Seleção univariada (Filter Methods):** Utiliza-se o coeficiente de correlação da feature com a variável target.



**- Vantagens:** Método simples, Método rápido e Fácil entendimento.

**-** **Desvantagens:** Não considera a influência entre as variáveis preditoras. Por exemplo, imagine que determinada variável tenha correlação fraca com a variável resposta, mas ao ser combinada com outra variável pode apresentar uma correlação forte. Essa técnica pode ser utilizada no EDA, mas para seleção de features para o modelo pode ser um pouco desastroso.

1. **Seleção por importância (Embedded Methods):** O método de seleção de features por importância é um subproduto do modelo.

**Ex:** Random Forest, Lasso Regression (algoritmo Regularizado), Ridge Regression.

**Random Forest:** Selecionar as variáveis que separam o conjunto de dados originais em dois outros de forma homogênea. O Gini impurity calcula a homogeneidade

1. **Seleção por subset (Wrapper Methods):**

## Fine Tuning

* Encontrar o conjunto de parâmetros que maximiza o aprendizado do modelo.
* Geralmente, não ocorre uma melhoria nas métricas utilizadas para a performance do modelo. O que realmente melhora a performance de um modelo é o acréscimo de mais uma variável que descreva o comportamento do fenômeno.

### 3 Estratégias para Fine Tuning:

1. Random Search:
2. Define valores para cada um dos hiperparâmetros aleatoriamente.
3. Rápido e fácil de implementar (Ponto positivo)
4. Pode repetir os mesmos hiperparâmetros nos testes (ponto negativo)
5. Grid Search:
6. Define todas as combinações possíveis de valores que os hiperparâmetros podem assumir.
7. Muito provável que essa estratégia realmente encontre o conjunto de parâmetros que favoreçam a melhor performance do modelo. (Ponto positivo)
8. Muito tempo de processamento (+custos) e pouca relevância na melhora da acurácia. (Ponto negativo)
9. Bayesian Search:
10. Define os valores para os hiperparâmetros seguindo a teoria de Bayes.
11. Define os valores dos parâmetros por meio de aprendizado (Ponto positivo).
12. É uma estratégia mais complexa que as outras (Ponto negativo)

