Introdução ao Python

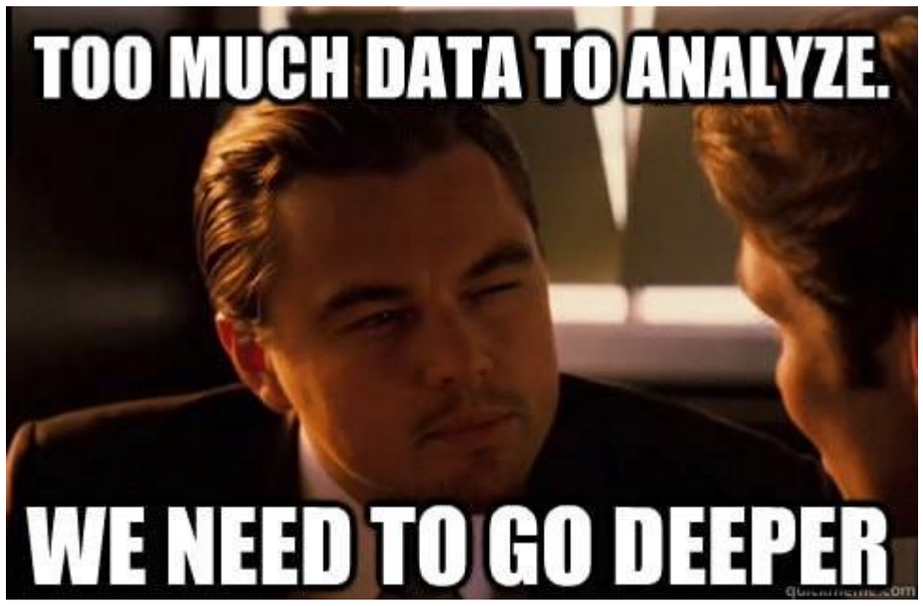
4. Análise Exploratória de Dados

Rodrigo Barbosa de Santis

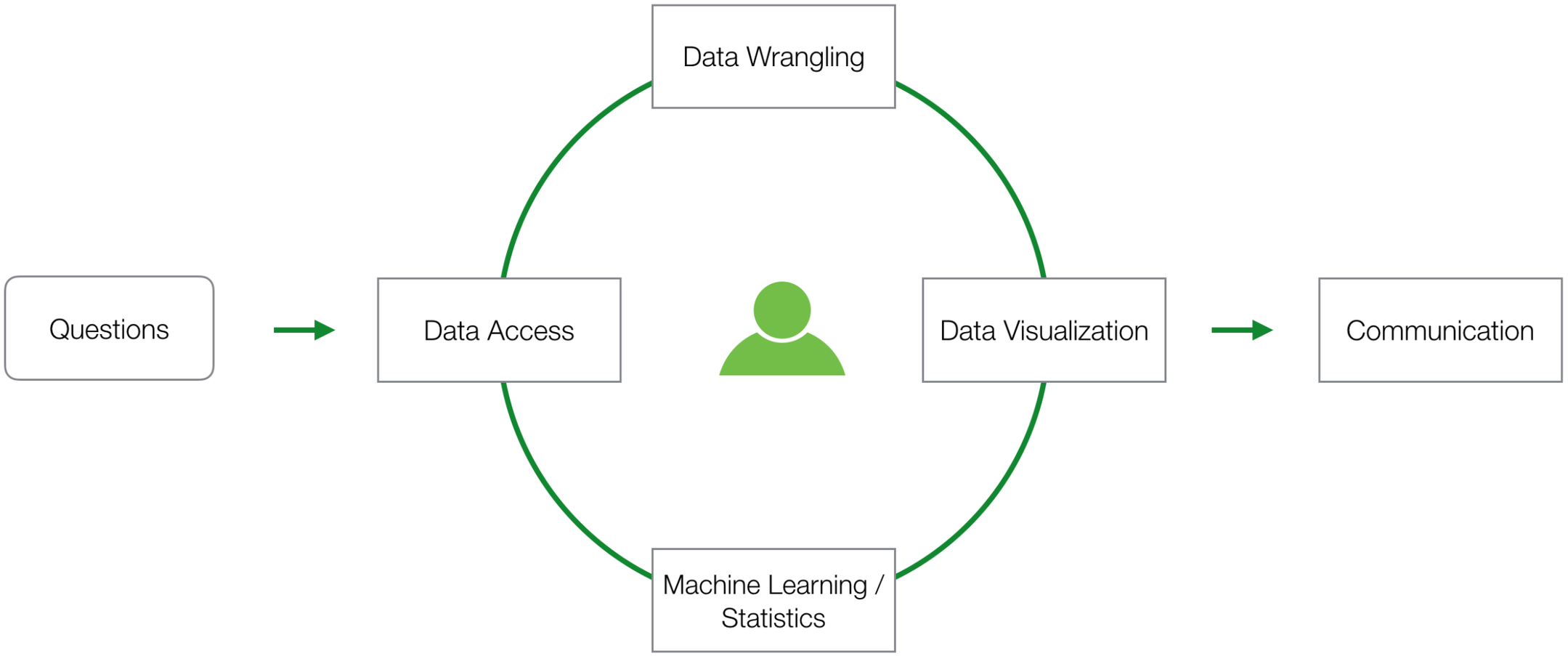
Sumário

Introdução

- Análise Exploratória de Dados ou Exploratory Data Analysis (EDA)
  - Embora nem sempre muito valorizado, é um dos componentes mais importantes para qualquer experimento de ciência dos dados.
  - Seu objetivo **entender** e **resumir** o conteúdo do conjunto de dados para garantir que os atributos que estamos alimentando nossos modelos de aprendizado de máquina são refinados, e para que obtenhamos resultados **válidos** e **interpretáveis**.



Data Science Workflow



- **Limpeza:** verificar problemas com os dados coletados, como dados ausentes ou erro de medição, tipo de dados das colunas etc.)

- **Perguntas:** definir perguntas, identificando relacionamentos entre variáveis que são particularmente interessantes ou inesperadas. Ex.: Existe um

- Usar visualizações eficazes para comunicar meus resultados (próximo slide).

Métodos

- Embora seja um campo muito aberto, em geral os métodos aplicados são divididos nos quatro grupos abaixo:

  - **Visualização univariada** - fornece estatísticas resumidas para cada atributo no conjunto de dados brutos;

  - **Visualização bivariada** - é realizada para encontrar o relacionamento entre cada variável no conjunto de dados e a variável de interesse;

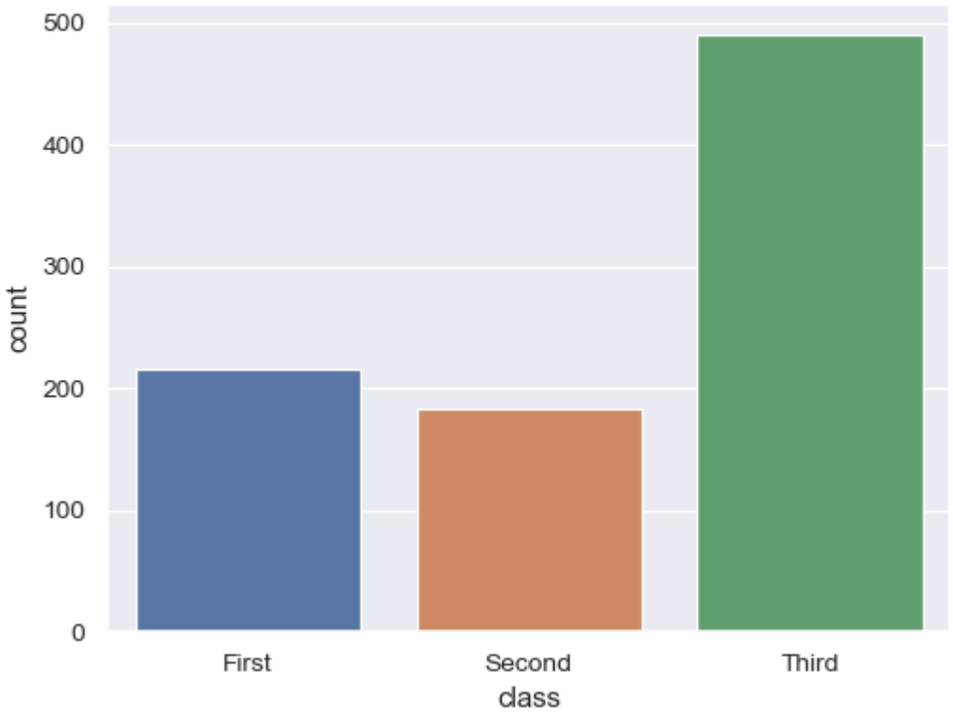
  - **Visualização multivariada** - é realizada para entender as interações entre diferentes atributos no conjunto de dados;

  - **Redução de dimensionalidade** - ajuda a entender os campos nos dados que representam a maior variação entre as observações e permitem o processamento de um volume reduzido de dados. (Veremos na parte de Engenharia de Atributos)

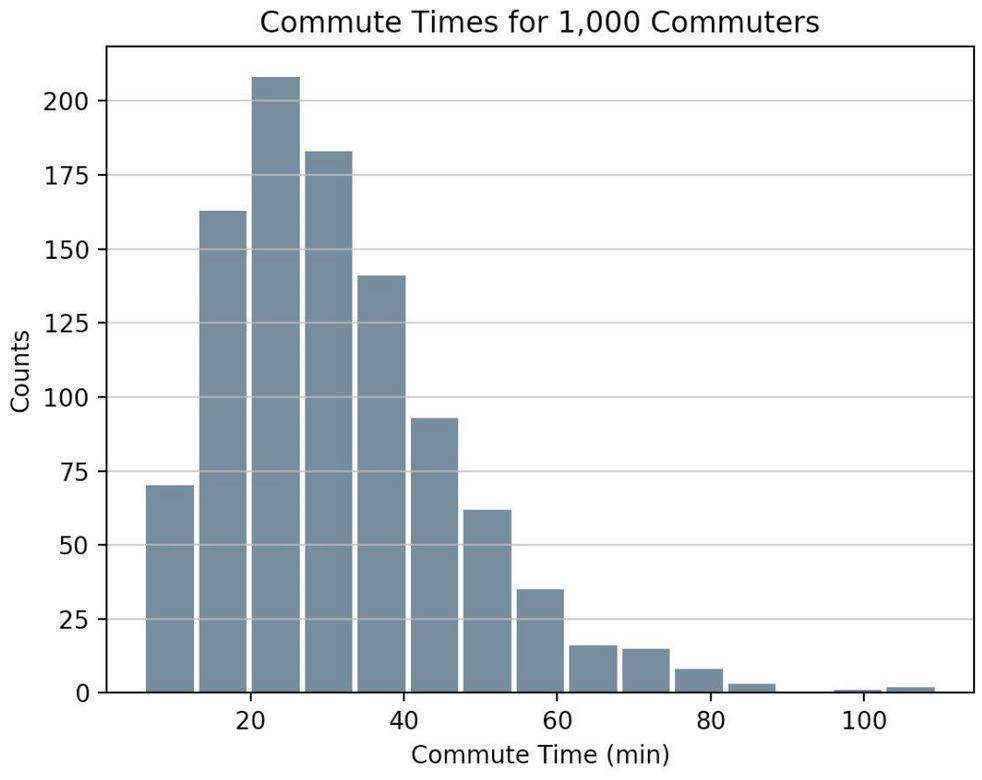
Variáveis numéricas x categóricas

- Estratégias mais comuns:
  - Variáveis **numéricas** com **histogramas**;
  - Variáveis **categóricas** com **gráficos de contagem**;
  - Relações entre variáveis numéricas com **gráficos de dispersão**, **gráficos conjuntos** e **gráficos de pares**;
  - Relações entre variáveis numéricas e categóricas com **gráficos de caixa** e **gráficos condicionais** complexos.

Visualização univariada

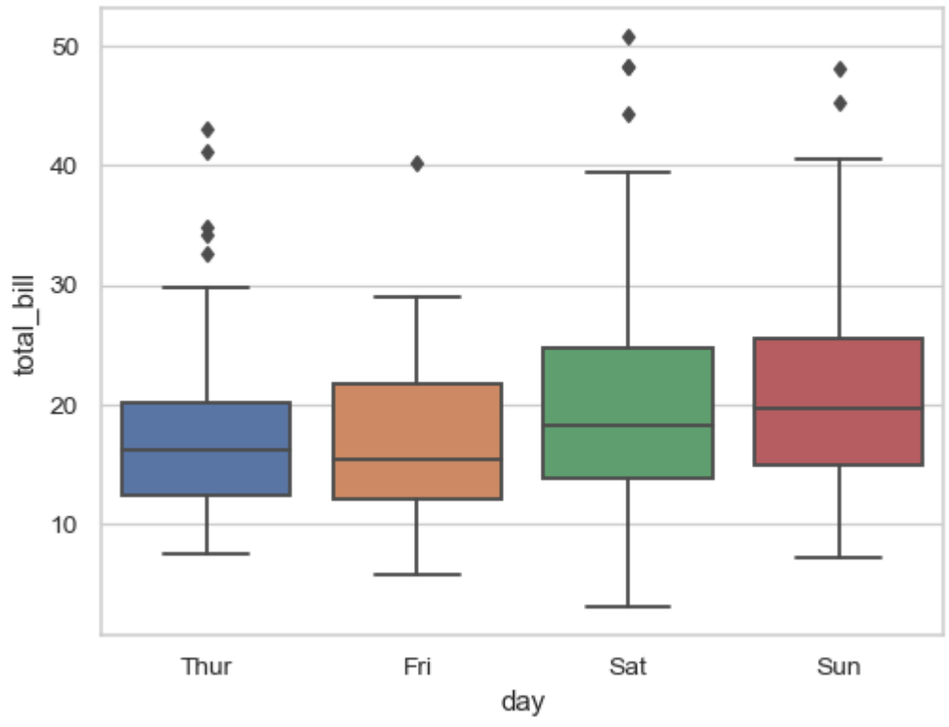


Categóricas: **Gráfico de Contagem**

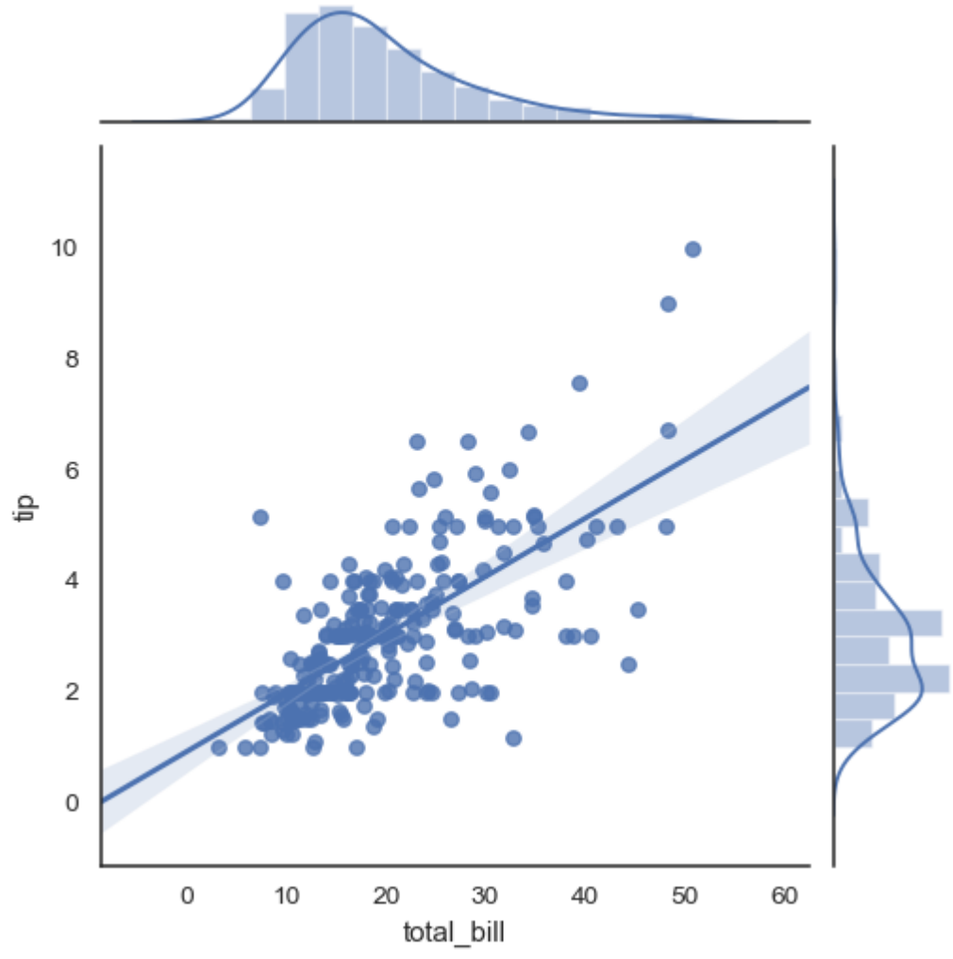


Categóricas: **Histogramas**

Visualização bivariada

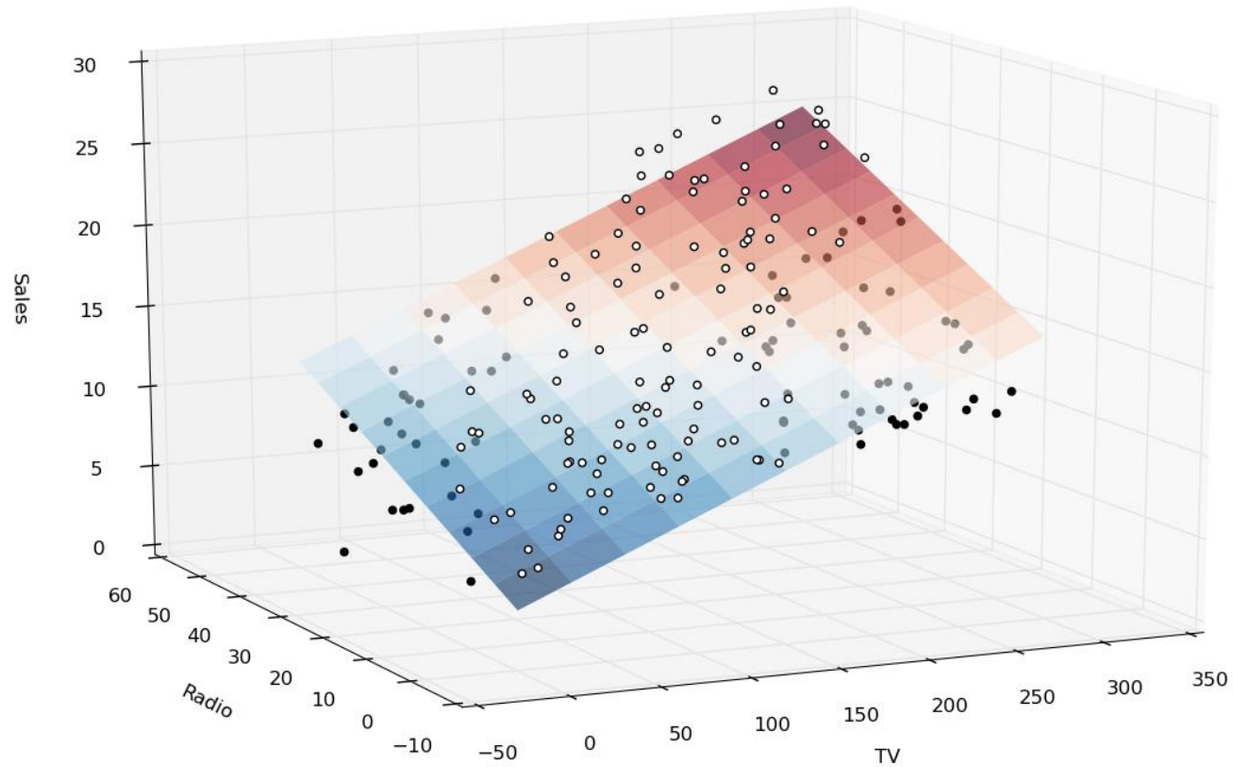


Categóricas: **Gráfico de Caixa**



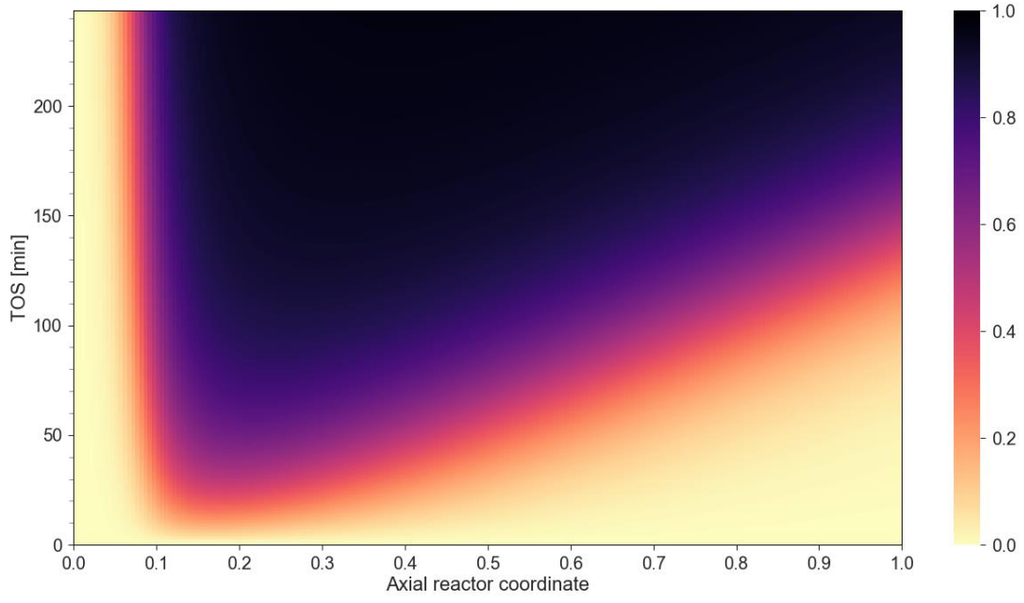
Numéricas: **Gráfico de Dispersões**

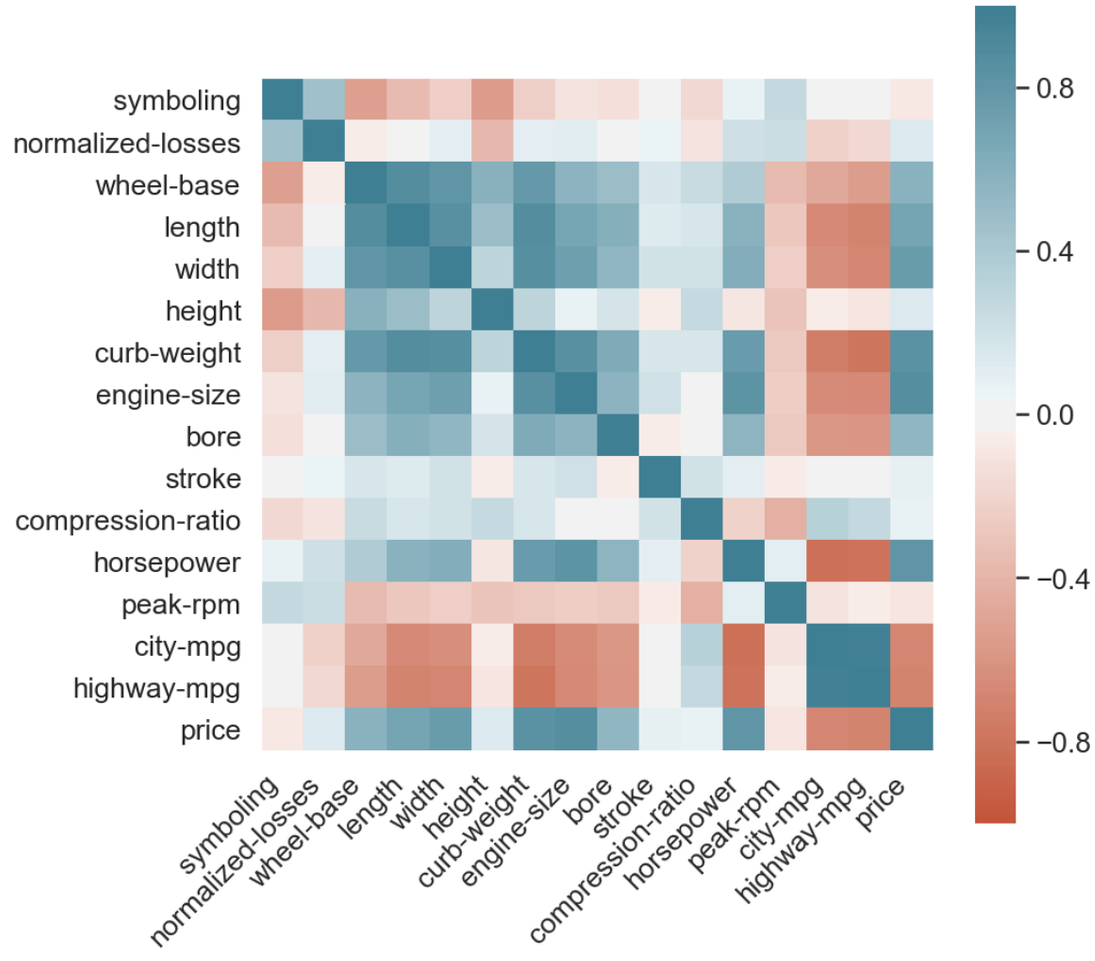
Visualização multivariada



Numérica: **Gráfico de Dispersão/Curvas 3D**

Visualização multivariada





Misto: **Mapas de Calor**

Conclusões

- EDA é uma arte: existe muita subjetividade e conhecimento sobre o fenômeno gerador dos dados em si.

- Não existe resposta certa, o jeito é meter a mão na massa e ver que insights.

- Embora subestimada, a EDA é muitas vezes tão ou até mais importante quanto a estimação do modelo em si.

- A prática leva a perfeição: quanto mais análises você fizer e ver de outras pessoas, melhor suas EDAs serão.

Exercícios

- Com seu grupo, escolha um banco de dados e faça uma análise exploratória EDA, realizando limpeza dos dados (se preciso), e formulando perguntas e respondendo-as a partir do uso das ferramentas de visualização.