Insper

Ciência dos Dados

Aula 02

Pandas e Seleção de Variáveis

Objetivos de Aprendizagem

Os alunos devem ser capazes de:

- Compreender as etapas envolvidas numa Análise Exploratória de dados e suas importâncias; e saber aplicá-las no Projeto 1.

- Manipular uma base de dados (limpeza, criação de novas variáveis, selecionando linhas e/ou colunas,...) , considerado a biblioteca PANDAS.

PROJETO 1 - PNAD

No site do IBGE, há dois formatos de *dataset*: arquivo de Pessoas – em que cada linha descreve as características das pessoas residentes em um domicílio.

O **Projeto 1 irá trabalhar com a PNAD PESSOAS** e tem como proposta levantar possíveis *semelhanças*, *diferença e/ou melhorias* em **ALGUMAS possíveis vertentes**:

- Descrever os **brasileiros quanto ao acesso à Internet e posse de telefone móvel celular para uso pessoal**; ou
- Descrever **trabalho infantil no Brasil**; ou
- Descrever **escolaridade no Brasil**; ou
- Descrever **salário dos brasileiros**; ou
- Descrever **outro perfil desde que antes discutido com a professora**.

Blackboard ou Github para ter acesso ao Projeto 1.

Análise Exploratória de Dados

A **Análise Exploratória de Dados** é uma das partes fundamentais do processo de Ciência dos dados.

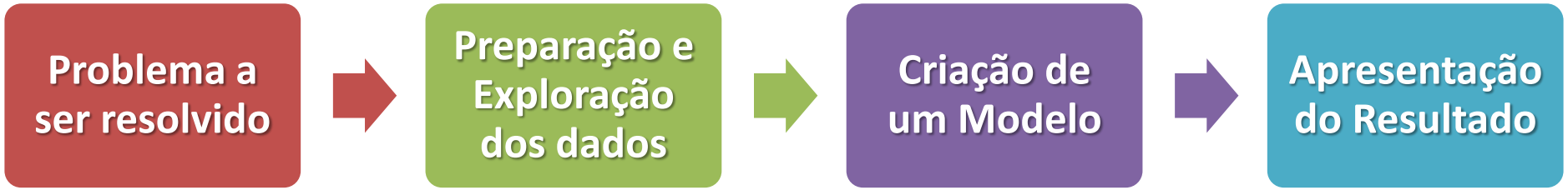
Entretanto, **"A qualidade dos *outputs* dependerá da qualidade dos *inputs*"**.

Ainda, mesmo que haja interesse em *Machine Learning*, por exemplo, antes será necessário deixar os dados organizados e consistentes. E isso faz parte da Análise Exploratória de dados.

Fonte: https://www.datascienceacademy.com.br/

Análise Exploratória de Dados

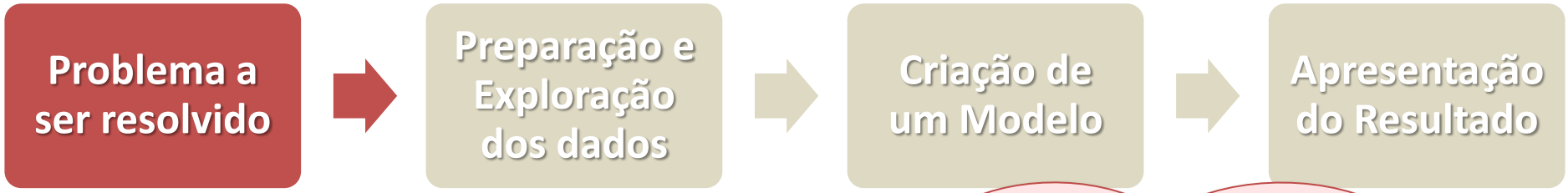
Etapas necessárias na **Análise Exploratória de Dados**:



Fonte: https://www.datascienceacademy.com.br/

Análise Exploratória de Dados

Etapas necessárias na **Análise Exploratória de Dados**:



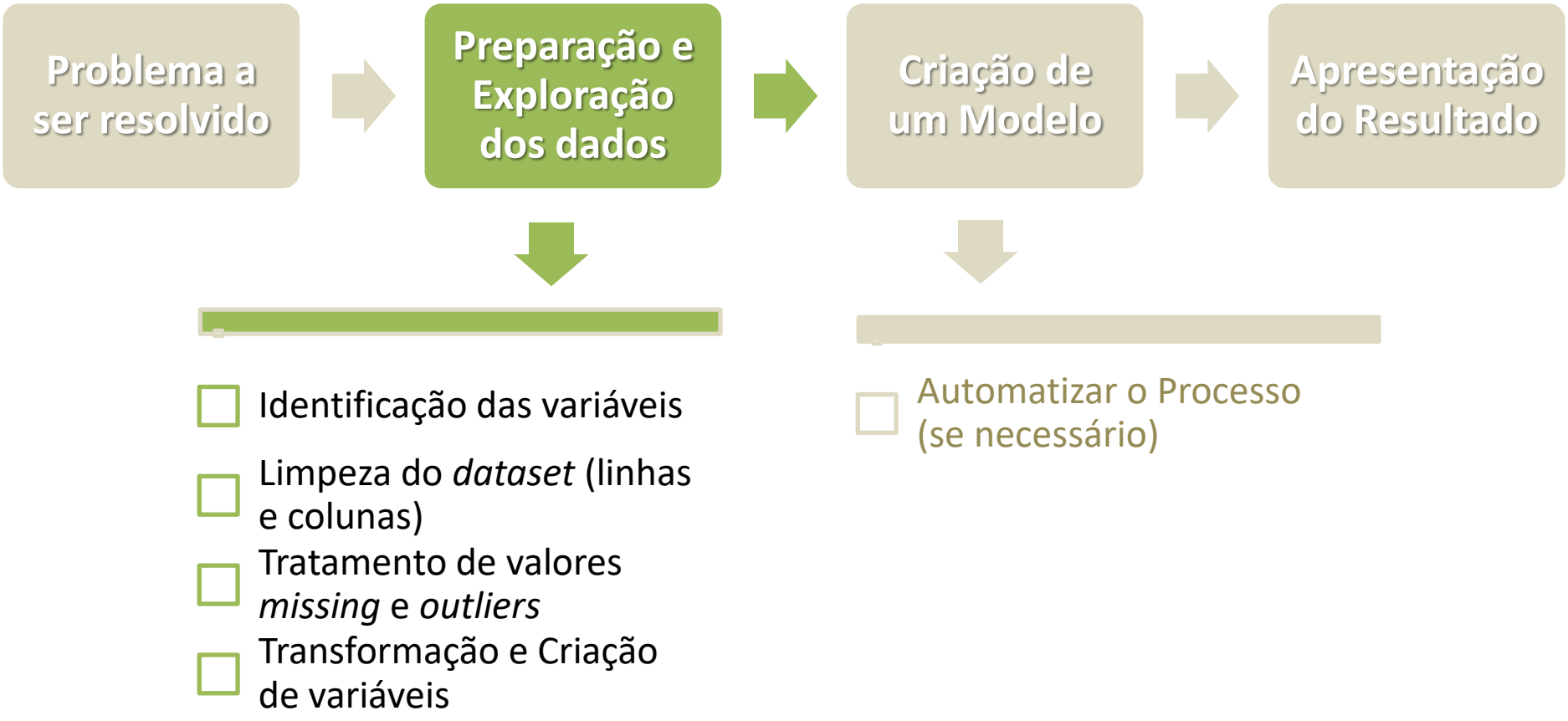
Um objetivo bem definido certamente auxilia nas escolhas das variáveis e norteia na escolha de um modelo.

Essa fase inicial é de muita importância, pois escolhas erradas aqui podem arruinar todas as demais etapas da análise de dados.

Fonte: https://www.datascienceacademy.com.br/

Análise Exploratória de Dados

Etapas necessárias na **Análise Exploratória de Dados**:



Fonte: https://www.datascienceacademy.com.br/

Análise Exploratória de Dados

Etapas necessárias na **Análise Exploratória de Dados**:

Consome de 60% a 70% de todo o processo.

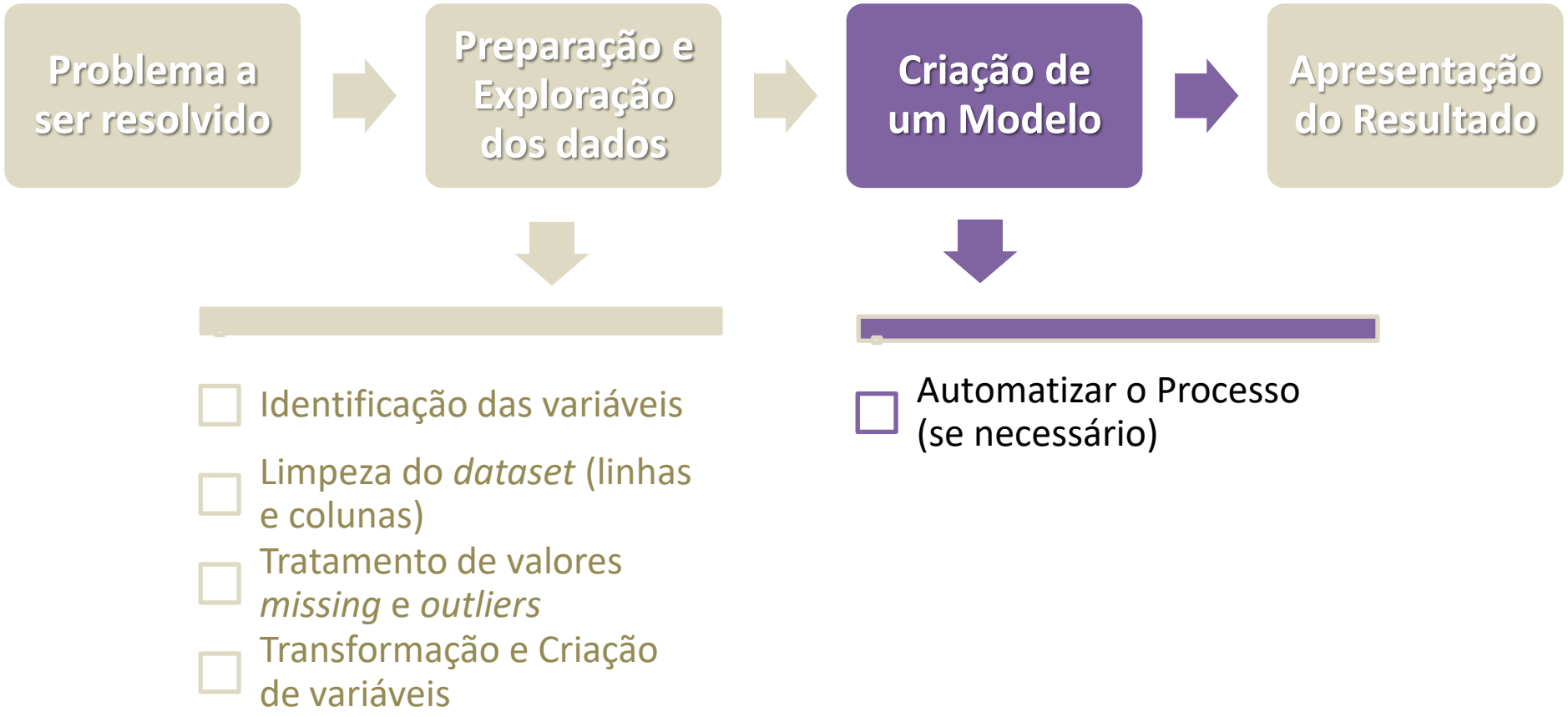
Só após ter a base de dados limpa e organizada (linhas e colunas do ***dataset*** **bem definidas), se faz a exploração dos dados SEMPRE NORTEADOS pelo problema.**

Fonte: https://www.datas[illegible]

Insper

Análise Exploratória de Dados

Etapas necessárias na **Análise Exploratória de Dados**:



Fonte: https://www.datascienceacademy.com.br/

Insper

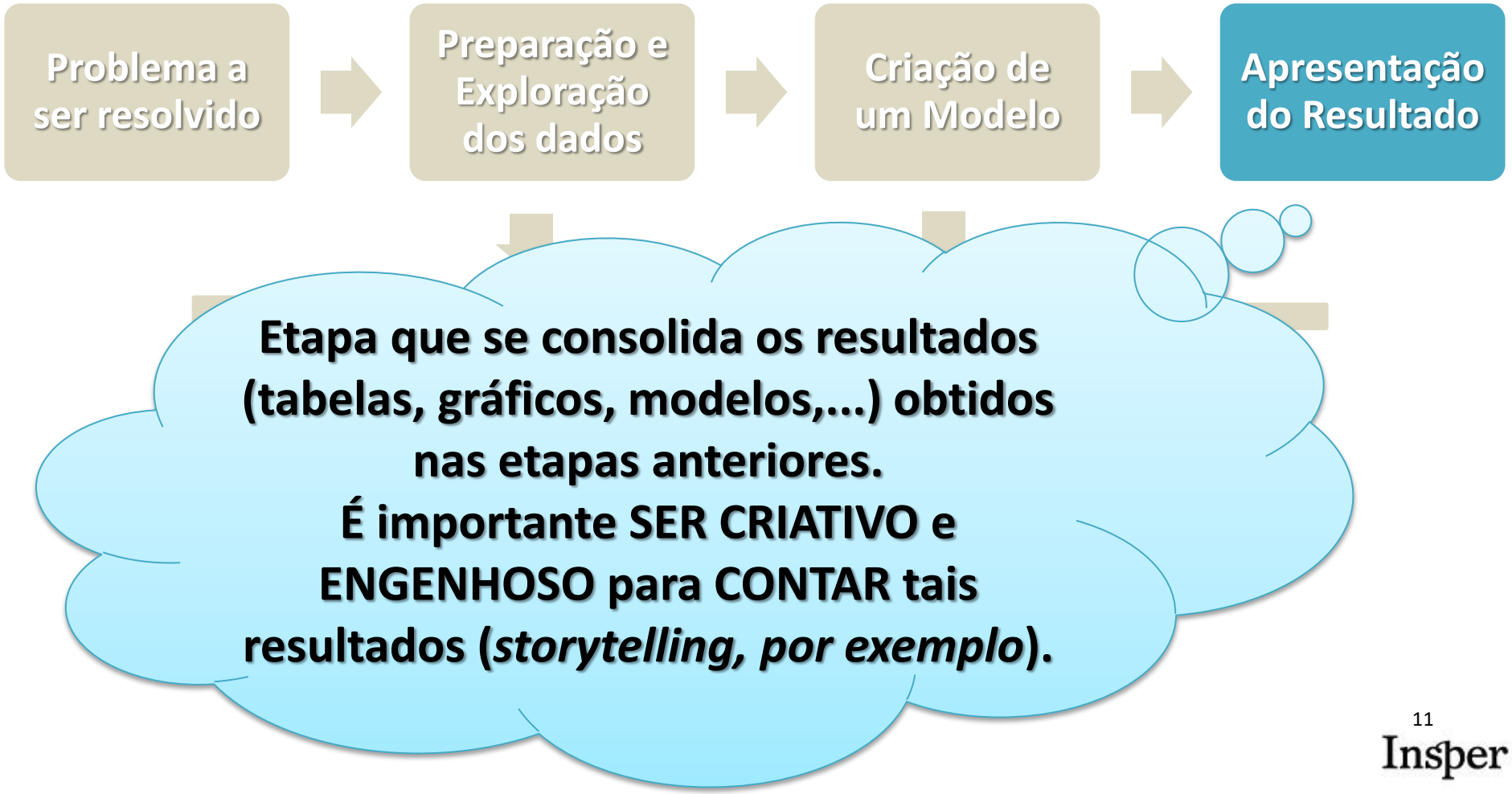
Análise Exploratória de Dados

Etapas necessárias na **Análise Exploratória de Dados**:

Fonte: https://www.datascienceacademy.com.br/

Análise Exploratória de Dados

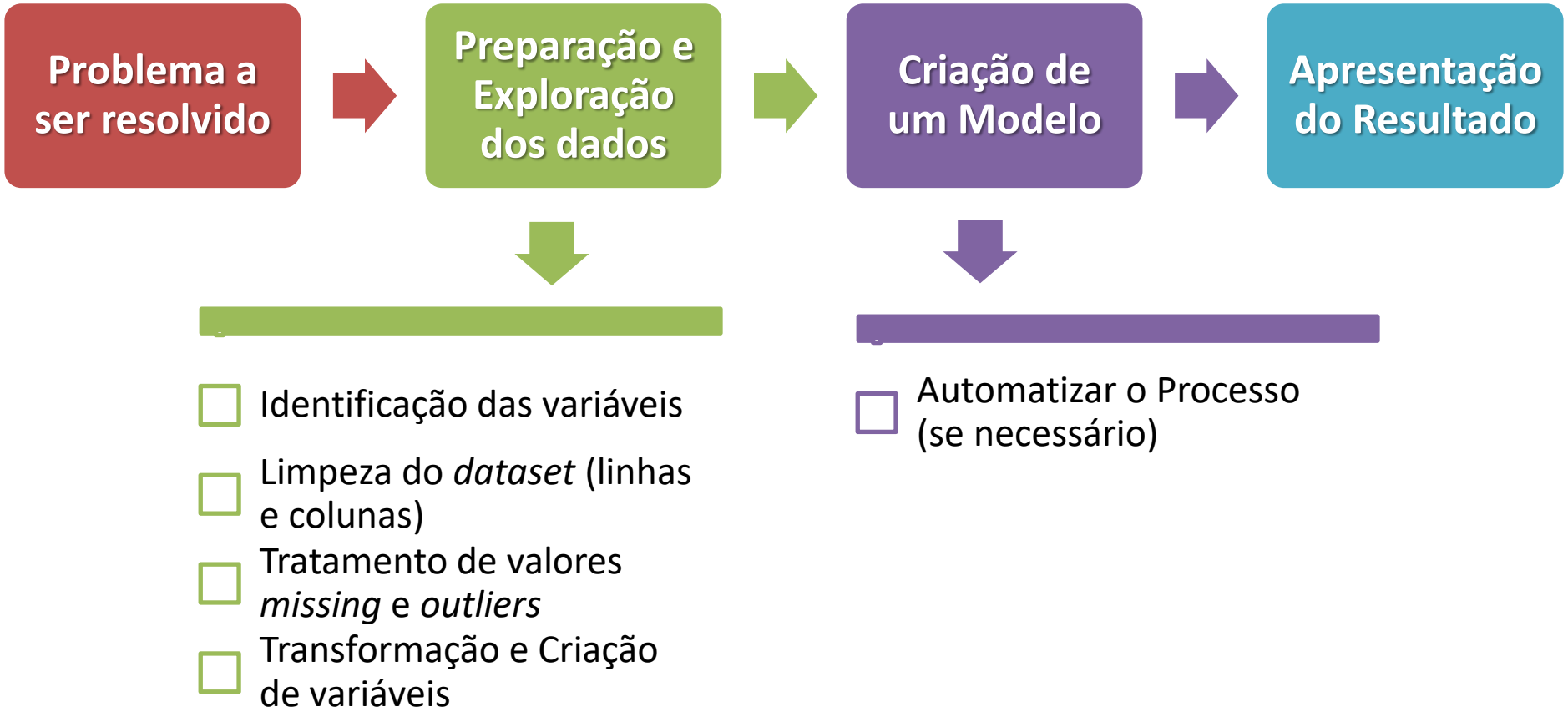
Etapas necessárias na **Análise Exploratória de Dados**:



Insper

Análise Exploratória de Dados

Etapas necessárias na **Análise Exploratória de Dados**:



Fonte: https://www.datascienceacademy.com.br/

Empresa de TV

PROBLEMA:

- Uma empresa de TV via satélite criou recentemente dois tipos de planos de canais (A e B).
- A empresa tem como objetivo:
  - Estudar o perfil dos clientes que aderiram cada plano para enviar malas diretas aos potenciais clientes de cada tipo de plano.

Usar base de dados `EmpresaTv Cod.xlsx`

Empresa de TV

Essa base de dados apresenta algumas variáveis para uma amostra de 82 clientes selecionados aleatoriamente dentre aqueles que aderiram aos planos.

As variáveis têm os seguintes significados:

*CLIENTE: identificador do cliente.

*PLANO: apresenta o plano adquirido pelo cliente – (1=A ou 2=B).

*EC: apresenta estado civil do cliente no momento da adesão ao plano – (1=Casado, 2=Solteiro e 3=Outros).

*SATISFACAO: grau de satisfação do cliente pelo plano –

(5=Muito satisfeito, 4=Satisfeito, 3=Indiferente, 2=Insatisfeito e 1=Muito insatisfeito).

*RENDA: renda pessoal do cliente, em milhares de reais.

Explorando a base de dados

Pelo Blackboard ou pelo Github, trabalhe com o arquivo:

`Aula02 Análise Exploratória Pandas e Seleção de Variáveis.ipynb`

PROJETO 1 - PNAD

No site do IBGE, há dois formatos de *dataset*: arquivo de Pessoas – em que cada linha descreve as características das pessoas residentes em um domicílio.

O **Projeto 1 irá trabalhar com a PNAD PESSOAS** e tem como proposta levantar possíveis *semelhanças*, *diferença e/ou melhorias* em **ALGUMAS possíveis vertentes**:

- Descrever os **brasileiros quanto ao acesso à Internet e posse de telefone móvel celular para uso pessoal**; ou
- Descrever **trabalho infantil no Brasil**; ou
- Descrever **escolaridade no Brasil**; ou
- Descrever **salário dos brasileiros**; ou
- Descrever **outro perfil desde que antes discutido com a professora**.

Blackboard ou Github para ter acesso ao Projeto 1.

APS 1 – Check durante próxima semana.

Devem apresentar aos ninjas:

1. Criar **NOVO** repositório no Github para CD!
2. Ter problema (OBJETIVO) definido!
3. Ter lido dataset original (para pelo menos um ano)
4. Ter uma versão salva com variáveis de interesse (colunas) e pessoas (linhas)!

O horário de CHECK será SEGUNDA das 18h às 19h30!! Deverão chegar ATÉ 20 minutos após início do horário de Check.

Preparo para próxima aula

Os alunos devem se preparar com:

1. Leitura prévia necessária: Magalhães e Lima (7ª. Edição): pág. 9 a 16 – destacando para variáveis quantitativas.
2. Projeto 1 – venham com objetivo definido e seleção das variáveis para a vertente escolhida.