© Al2 – Advanced Institute for Artificial Intelligence

O conteúdo do **Programa de Residência em IA**é de propriedade exclusiva do Al2 sendo cedido
para uso, único e exclusivo, do(a) aluno(a),
não podendo ser compartilhado, distribuído,
comercializado e/ou gravado, seja da forma que for.



# Programação em Python

https://advancedinstitute.ai



## Programação em Python

Pandas e dados Missing

#### Referências

### Referências e Fontes das Imagens

- □ Pandas Oficial
- □ Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (Book)
- ☐ Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visualization (Book)

## Aquisição dos Dados

### Introdução

- □ Dados faltantes são um sério problema quando se trata de análise de dados.
- □ Essa incompletude pode ser derivada de diversas falhas no armazenamento da informação, desde perca de dados em sistema até falhas humanas de preenchimento.
- ☐ A falta de dados nas amostras cria lacunas no DataSet, que por consequência cria alterações nas informações que se busca obter.
- □ Uma análise incompleta pode causar incerteza nos resultados, e por consequência, impactar a tomada de decisões

## Aquisição dos Dados

#### Introdução

- □ A análise de dados demanda precaução durante todo o seu processo.
- □ Depois da coleta de dados, dados faltantes provavelmente estarão presentes, o que é muito comum em diversos tipos de bancos de dados, especialmente em dados epidemiológicos.
- ☐ Existem diversos métodos para isso, que são divididos em duas categorias:
  - Métodos de Imputação Simples: que atribuem um único valor para cada dado faltante;
  - Métodos de Imputação Múltipla: que atribuem diversos valores para o mesmo

## Imputação Simples X Imputação Múltipla

## Método Ingênuo (Imputação Simples)

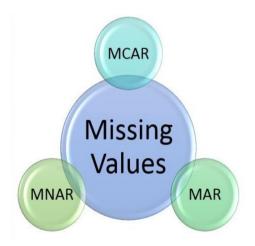
- ☐ Leva somente em consideração informações da variável que contém dados faltantes;
- □ Completa os "espaços em branco" através de cálculos de estatística descritiva;
- 🗆 Utilizado quando o dado faltante diz respeito a uma variável contínua.
- **Média:** Cada valor faltante na variável *Y*, será preenchido com a média dos valores observados da variável *Y*.
  - Moda: Método utilizado quando o dado faltante se diz respeito a uma variável binária.
- Regressão: Busca definir uma relação, através de uma equação, entra a média de uma variável aleatória  $y_i$  chamada de variável resposta, e outras variáveis conhecidas  $x_{ij}$ , j=1,...,p, chamadas de variáveis explicativas.

## Imputação Simples X Imputação Múltipla

## Método Imputação Múltipla

- ☐ Método utilizado tanto quando o dado faltante diz respeito a uma variável contínua quanto a binária;
- □ Deve assumir que os dados faltantes sigam o mecanismo MAR ou mesmo MNAR porém, para este último necessita aplicar alguns ajustes.

Uma das questões importantes com dados ausentes é o mecanismo de dados ausentes.



## Existem 3 principais tipos de dados faltantes:

- ☐ Missing Completely at Random (MCAR): ocorre quando o valor faltante não depende dos dados observados e nem dos não observados; é um evento aleatório.
  - ex: sujeito pula uma questão sem querer



## Missing Completely at Random (MCAR)

- □ Consequências tendem a ser pequenas, já que o padrão de *missing* não tem relação com as informações (são ao acaso)
  - tende a n\u00e3o afetar um item mais que os outros;
  - tende a não influenciar muito no padrão de respostas do questionário;
- □ Por ser completamente aleatório, tende a representar um baixo percentual do total de respostas do BD, principalmente sobre aquele item.

#### Existem 3 principais tipos de dados faltantes:

- ☐ **Missing at random (MAR):** acontece quando a ausência de respostas não tem relação direta com a questão, pois é causada por um agente externo ao item.
  - ex: homens provavelmente respondam sobre o seu peso mais do que mulheres, logo, a variável peso é MAR.



## Missing at random (MAR)

- O dado faltante depende dos valores observados, ou seja, uma variável que contem os dados faltantes depende de uma variável com dados observados. Logo, a falta se refere a uma variável particular.
  - possível controlar o efeito da variável preditora do missing.

#### Existem 3 principais tipos de dados faltantes:

- Missing not at random (MNAR): aparece em situações onde existe uma razão especifica para o dado faltante, ou seja, está relacionado aos valores não observados.
  - ex: pessoas com depressão talvez rejeitem preencher uma pesquisa sobre depressão



## Missing not at random (MNAR)

 Muito comum quando pessoas não querem responder algo pessoal ou seja, a ausência de resposta está ligada diretamente ao item.

## Resumo dos 3 principais tipos de dados faltantes:

	MCAR	MAR	MNAR
Variável (Item)	Sujeitos omitem respostas aleatoriamente	Sujeitos omitem respostas que podem ser consegui- das por outras variáveis	Sujeitos não respondem itens sem algum tipo de critério
Indivíduos ou sujeitos	Faltam dados de sujeitos aleatoriamente	Faltam dados de sujei- tos, mas que são relacio- nados com os dados de- mográficos disponíveis	Faltam dados de sujeitos e são relacionados com os dados demográficos não medidos
Ocasiões	Sujeitos aleatoriamente não se apresentam na sessão	Sujeitos que se desempe- nham mal na sessão ante- rior, não se apresentam na sessão seguinte	Sujeitos que estão se desempenhando mal na sessão atual, deixam de participar

Dúvidas?