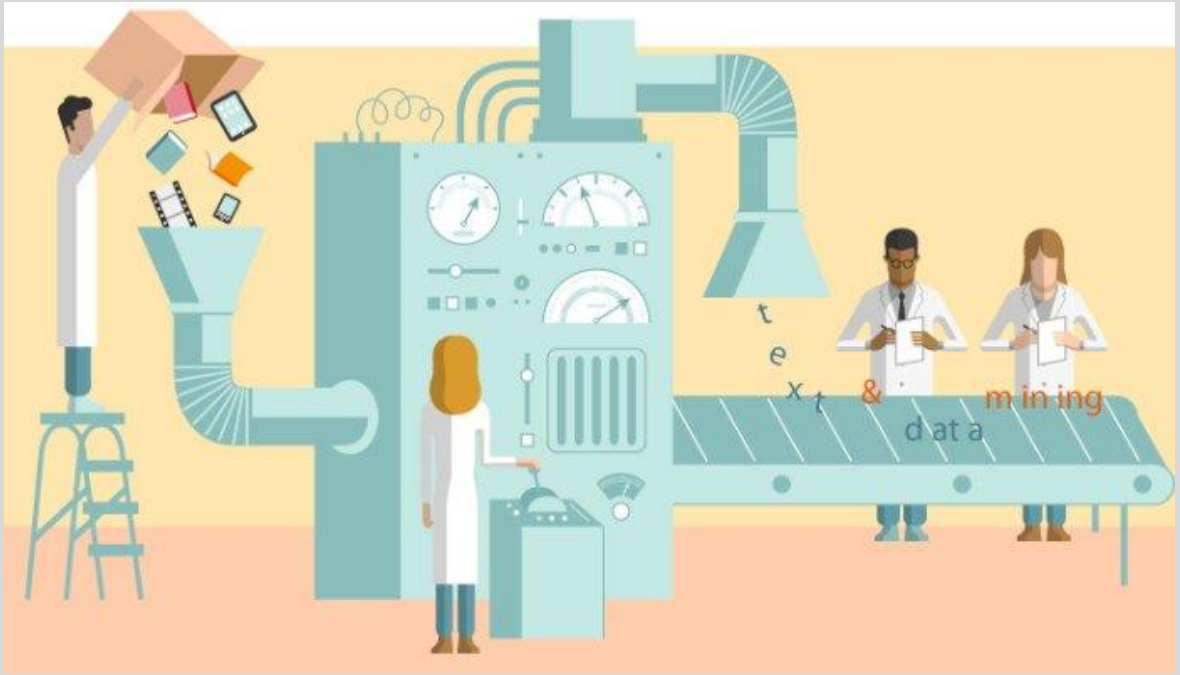
1. Ajuda o Edu: coloca o seu nome no Zoom para eu poder fazer a presença de forma mais fácil

2. Com as aulas remotas, perdemos uma coisa legal da aula presencial: as conversas antes da aula. Nessas conversas nos conhecemos melhor.

3. Minha sugestão para quem entrar antes, é trocar ideias sobre as dúvidas das aulas ou usar alguns ‘quebra-gelo’ tradicionais:
   - Onde trabalha? O que faz?



DigitalHouse > Coding School

DATA SCIENCE

Introdução ao Pandas

Julho 2020

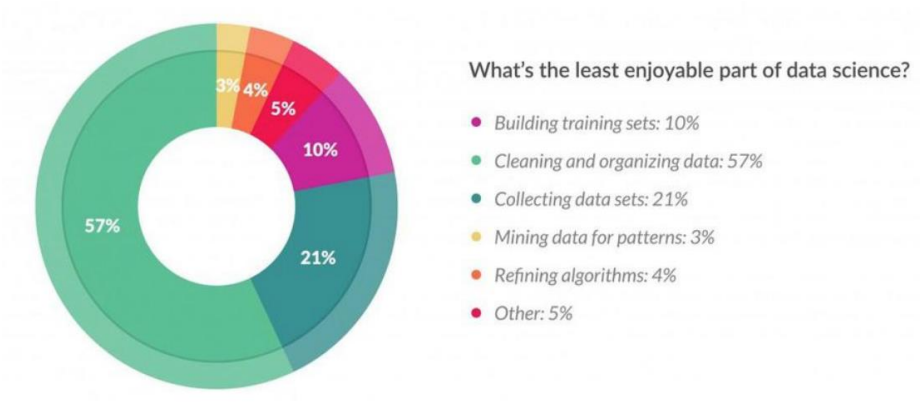
INTRODUÇÃO AO PANDAS

1. Limpeza e preparação de dados
2. O que é o Pandas?
3. Para que serve o Pandas?

Limpeza e preparação de dados

Limpeza e preparação de dados

— Em 2009, Mike Driscoll (cientista de dados e CEO da Metamarkets) popularizou o termo **“data munging”** em referência ao **trabalhoso processo de limpar, preparar e validar os dados**



Fonte: Forbes.

Limpeza e preparação de dados

- **Em 2013, Josh Wills** (ex-diretor de Data Science da Cloudera e atual Diretor de Engenharia de Dados na Slack) comentou: "**I'm a data janitor.** That's the sexiest job of the 21st century. It's very flattering, but it's also a little baffling."

Tradução:

Em Data Science, 80% do tempo é investido na preparação dos dados e os 20% restantes em reclamar da necessidade de preparar os dados

Python vs. outros para a limpeza de dados

- Processo reproduzível.
- Controle de versões.
- Criação de testes automáticos.
- Facilita a manutenção.
- Linguagem dinâmica que permite uma alta produtividade.
- Possibilidade de criar interface com C para melhorar o desempenho.
- Preferido em ambientes de data Science.
- You shouldn't use a spreadsheet for important work (I mean it).

Pandas

Pandas

- **Pandas é uma coleção de funções e estruturas de dados** que facilitam o trabalho com dados estruturados.
- Inicialmente construído **com base em Numpy** por Wes McKinney.
- Nome derivado de "**Pan**el **Da**ta **S**ystem" (termo econométrico para conjuntos de dados multidimensionais).
- Fornece recursos de manipulação de dados flexíveis semelhantes a spreadsheets e bancos de dados relacionais.
- Combina o alto desempenho das **operações sobre arranjos de NumPy com a flexibilidade na manipulação de dados,** de uma spreadsheet ou um banco de dados relacional.
- Fornece **funcionalidades avançadas de indexação** para facilitar a manipulação, adição e seleção de partes de um conjunto de dados.
- Fornece **operações de agrupamento por colunas, filtros e sumarizações.**

Objetos no Pandas

- Veremos os seguintes objetos:
  - Series.
  - DataFrames.
  - Index.

Pandas

- No Pandas, duas estruturas de dados fundamentais são utilizadas: ***Series*** e ***DataFrames*****.**
- As duas estruturas usam arranjos* de **Numpy** como base.
- Uma *Series* é um arranjo unidimensional capaz de guardar qualquer tipo de dados (inteiros, strings, floats, objetos Python, etc.).
- Um *DataFrame* é uma matriz bidimensional. Pode ser visto como um conjunto de *Series* que compartilham os mesmos valores no índice.

* arrays

Pandas: Series

- Uma Series é um objeto semelhante a um **vetor unidimensional**.
- Contém **um array de valores e um array associado de tags** desses valores, denominado índice.
- Assim como os arrays de NumPy, permitem passar uma lista de valores com índices para selecionar um subconjunto de valores.

Pandas: Series

Pandas: DataFrame

- Representa uma **estrutura de dados tabular** que contém uma **coleção de colunas**, cada uma delas com um tipo determinado (number, string, boolean, etc.).
- Inspirados na estrutura data.frame de R.
- Permitem operações “ricas” sobre índices equivalentes aos comandos JOIN e GROUP BY em SQL.
- Ideais para organizar o resultado de uma análise em um formato útil para representá-lo graficamente ou exibi-lo.

Pandas: DataFrame

- O conceito de *dataframe* vem do mundo dos softwares estatísticos usados em pesquisa empírica. Geralmente se referem a dados tabulados.
- Uma estrutura de dados representando casos (linhas), cada um consistindo de um número medidas (colunas). Alternativamente, cada linha pode ser tratada como observações com múltiplas variáveis ou *features*.
- Os tipos de dados da linha (‘registro’) podem ser heterogêneos, enquanto que o tipo da coluna deve ser homogêneo.
- *Dataframes* usualmente contêm algum metadado em adição ao dado. Por exemplo, nome de coluna e de linhas.

Pandas: DataFrame

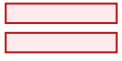
	Animais	Donos
0	Cachorro	João
1	Urso	Pedro
2	Girafa	Cristian
3	Tigre	Estêvão
4	Cobra	Pablo
5	Rato	Cláudio

Prática guiada 1
Objetos no Pandas

Indexação booleana no Pandas

	Nome	Sobrenome
0	João	da Silva
1	Pedro	Garcia
2	Mateus	Savala

True
False
True



	Nome	Sobrenome
0	João	da Silva
1	Mateus	Savala

Indexação e Subsetting no Pandas

- Para acessar os objetos no Pandas, é necessário introduzir os seguintes métodos:
  - .loc()
  - .iloc()

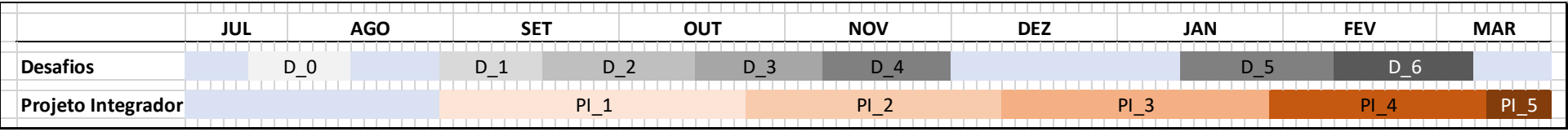
Prática guiada 2
Slicing e Indexing no Pandas

Conclusões

Pandas: DataFrame

- No dia a dia de um cientista de dados, a tarefa que mais consome tempo é a limpeza, preparação e normalização dos dados que serão trabalhados.
- Python é uma linguagem com numerosos recursos que podem facilitar essas tarefas.
- O Pandas é a principal biblioteca na hora de realizar essas tarefas.
- A partir de agora, os DataFrame do Pandas serão a ferramenta fundamental que vamos utilizar ao longo do curso e no nosso trabalho diário.

Desafios



Produto	Tema	Inicício	Entrega	Duração (d)
D_0	Data Wrangling	27/jul	12/ago	16
D_1	Estatística	31/ago	16/set	16
D_2	ML - Regressão	18/set	14/out	26
D_3	ML - Model Selection	16/out	06/nov	21
D_4	ML - Não Supervisionados	09/nov	30/nov	21
D_5	ML - Ensembles	13/jan	08/fev	26
D_6	Time series / RecSys	10/fev	05/mar	23
PI_1	Data Collection	31/ago	26/out	56
PI_2	EDA	26/out	09/dez	44
PI_3	Modelling	09/dez	27/jan	49
PI_4	Tunning	27/jan	08/mar	40
PI_5	Prez	08/mar	19/mar	11

Desafio 0