Obtenção de Dados

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

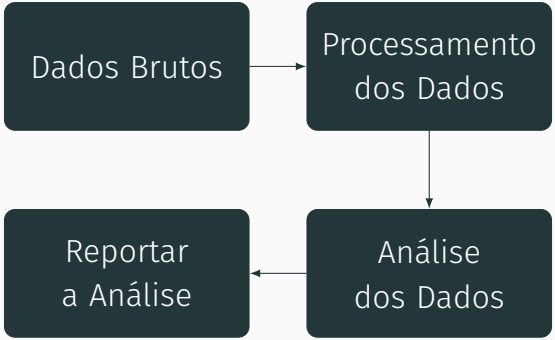
Universidade Federal do Rio Grande do Norte

Introdução

Introdução

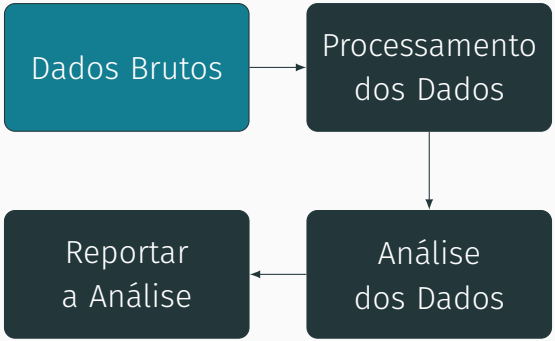
- Nem sempre é fácil conseguir dados para análise
- Como é possível extrair diversas informações a partir de análises, muitos vezes o acesso a dados é cobrado
- Veremos aqui como encontrar e baixar dados gratuitos da internet
- Além disso, vamos aplicar técnicas para o processamento destes dados visando sua análise posterior

Introdução


Dados Brutos
Processamento dos Dados
Análise dos Dados
Reportar a Análise

Dados Brutos

Dados Brutos



Dados Brutos

- Dados brutos são dados sem tratamento algum
- São os dados originais, vindos diretamente da fonte
- Através dos dados brutos conseguimos as tabelas que serão utilizadas em nossas análises

Dados Brutos

- Há diversas maneiras de dados brutos serem obtidos a partir da internet
- As maneiras principais são
  1. Download via sites de divulgação
  2. Web scraping
  3. Solicitação direta

Download de Dados Brutos

- Instituto Brasileiro de Geografia e Estatística (IBGE) - `https://downloads.ibge.gov.br/`
- Banco Central do Brasil - `https://www3.bcb.gov.br/sgspub/`
- Portal Brasileiro de Dados Abertos - `http://dados.gov.br`
- Portal da Transparência - `http://www.transparencia.gov.br/`
- UFRN - `http://dados.ufrn.br/`

Download de Dados Brutos

- Brasil.io - `https://brasil.io/`
- Kaggle - `https://www.kaggle.com/datasets`
- UCI Machine Learning Repository - `https://archive.ics.uci.edu/ml/`
- The home of the U.S. Government's open data - `https://www.data.gov/`

Download de Dados Brutos

- O R possui vários conjuntos de dados já pré-instalados para análise
- O comando

```
> data()
```

  lista todos os conjuntos de dados disponíveis nos pacotes carregados na memória do R
- O comando

```
> data(package = .packages(all.available = TRUE))
```

  lista todos os conjuntos de dados disponíveis em todos os pacotes do R disponíveis na sua instalação local

Exercícios

Exercícios

1. Vá ao Kaggle e baixe o conjunto de dados `Pokemon with stats`
2. Carregue este conjunto de dados no `R`
3. Conte as occorrências dos tipos de habilidades principais entre os pokémons (`Type.1`) e as classifique em ordem decrescente
4. Crie um boxplot com o poder de ataque (`Attack`) dos Pokemons e ordene o eixo x de acordo com a ordem decrescente das medianas para cada nível da variável `Type.1`
5. Identifique, através de um gráfico e uma regressão linear, se existe relação entre o valor de ataque e defesa dos pokémons (`Attack` e `Defense`)

Web Scraping

Web scraping

- Termo que significa vasculhar a internet através de dados
- A ideia é coletar e organizar automaticamente os dados que estão espalhados em um ou mais sites
- Logicamente, apenas dados abertos ao público podem ser coletados desta maneira

Web scraping

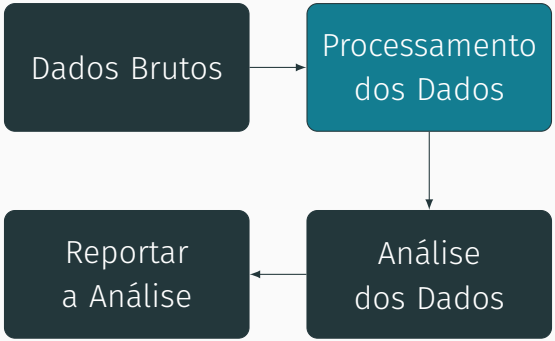
- O `R` possui pacotes que realizam este tipo de trabalho
- Existem desde pacotes bastante específicos, como o `twitteR`, até pacotes com usos mais gerais, como o `rvest` e `XML`
- O desempenho do `R` é bastante satisfatório para este tipo de situação

Web scraping

- Abra o arquivo `01_Obtencao_De_Dados_Codigos.R` para ver uma aplicação de web scraping utilizando o `R`
- Veremos como obter os dados de todos os pilotos da história da Fórmula 1 a partir da Wikipedia

Processamento dos Dados

Processamento dos Dados

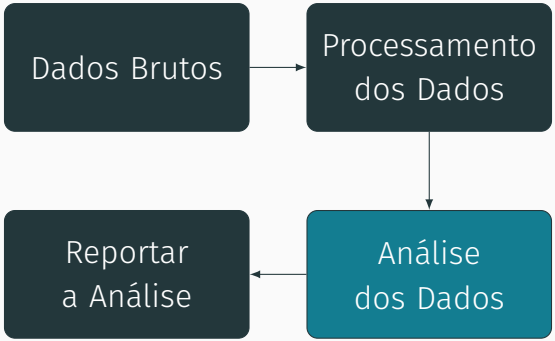


Processamento dos Dados

- Muitas vezes os dados brutos não estão prontos para análise
- Cada caso é um caso; técnicas que servem para preparar um conjunto de dados muito provavelmente não vão servir para preparar outro conjunto
- Vamos ver algumas técnicas aplicadas nos dados dos pilotos de Fórmula 1

Análise dos Dados

Análise dos Dados



Análise dos Dados

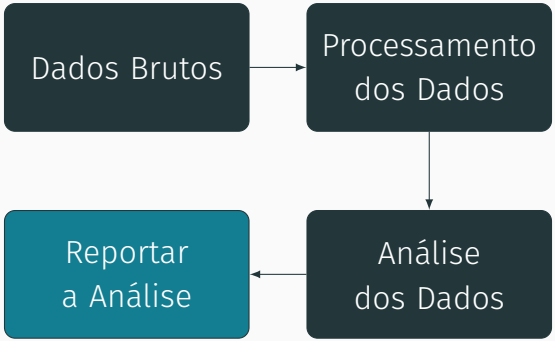
- Fica imensamente mais fácil proceder com a análise dos dados após eles terem sido previamente processados
- Perceba que o banco de dados criado com os dados dos pilotos de Fórmula 1, após ser processado, é muito mais fácil de analisar do que o conjunto original
- Ele está pronto para ser trabalhado por um programa como `R`, `SAS` ou até mesmo `Excel`

Exercícios

1. Classifique os pilotos Fórmula 1 em ordem decrescente de número de campeonatos, informando seu nome e país de origem
2. Encontre os 10 países com mais títulos na modalidade
3. Faça um gráfico do número de vitórias dos pilotos versus número de pole positions
4. Altere a cor dos pontos do gráfico anterior, identificando-os de acordo com o número de campeonatos vencidos por cada um

Reportar a Análise

Reportar a Análise



Reportar a Análise

- As conclusões tiradas na análise são reportadas neste passo
- Tente mantê-las interessantes e simples, além de concisas e completas
- Uma tendência atual é usar dados para contar uma história; tente se valer de uma abordagem assim para manter seu interlocutor interessado

Exercícios

Exercícios

1. Encontre e baixe no `R` a página da Wikipedia PT com a lista dos municípios brasileiros por população
2. Encontre e baixe no `R` a página da Wikipedia PT com a lista dos municípios brasileiros por área
3. Combine os resultados dos itens 1 e 2 em um data frame só, utilizando a função `left_join` do pacote `dplyr`
4. Limpe o data frame resultante do passo acima deixando-o apenas com as colunas `municipio`, `estado`, `area` e `populacao`, nesta ordem, sem acentos ou letras maiúsculas

Exercícios

5. Crie o gráfico de dispersão de população versus área. O que é possível perceber? Altere os eixos, um de cada vez e simultaneamente, para a escala logarítmica na base 10 e veja se a visualização melhora.
6. Encontre as cinco cidades com maior e menor população do Brasil
7. Encontre as cinco cidades com maior e menor densidade populacional do Brasil
8. Encontre as cinco cidades com maior e menor população do Rio Grande do Norte
9. Encontre as cinco cidades com maior e menor densidade populacional do Rio Grande do Norte

Obtenção de Dados

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte