Métodos de Estatística Aplicada com Python

Aula 1

Carlos Góes[1]

[1]Pós-Graduação em Ciência de Dados
Instituto de Educação Superior de Brasília

2017

Sumário

Sumário

Professor

Por que eu estou dando aula para vocês?

- Educacional
  - PhD em Economia (em andamento), UCSD
  - Mestre em Economia Internacional (2013), Johns Hopkins
  - Bacharel em Relações Internacionais (2011), UnB
- Profissional
  - Assessor Especial para Desenvolvimento Econômico, Secretaria Especial de Assuntos Estratégicos (2017-)
  - Pesquisador-Chefe, Instituto Mercado Popular (2016-)
  - Analista Econômico, FMI (2013-17)
  - Pesquisador, Instituto Cato (2012)

O que é ciência de dados?

O que faz você diferente de um estatístico ou um cientista da computação?

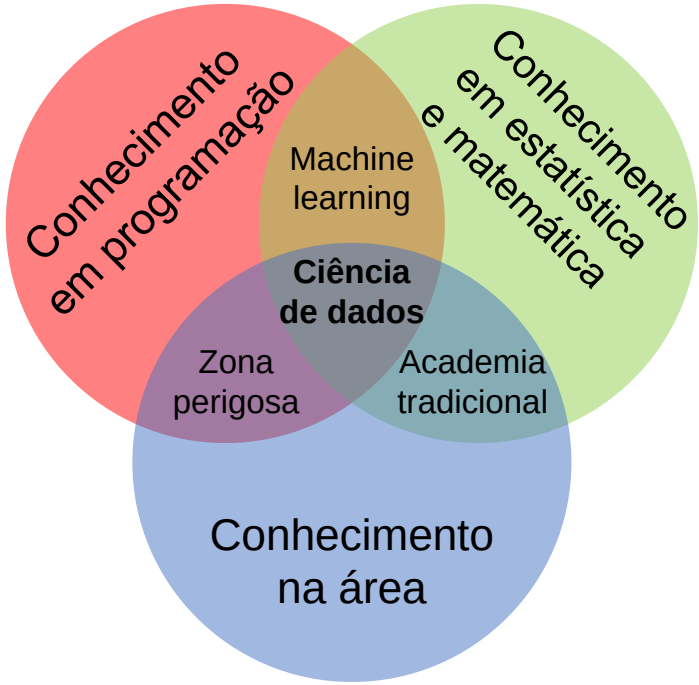

Figura: Diagrama de Conway. Componentes que compõem a ciência de dados e suas intersecções

Estrutura do curso

O que vamos fazer aprender aqui?

- O objetivo do curso é prover instrumentos téoricos e práticos necessários à compreensão de:
  - tipos distintos de variáveis;
  - apresentação de dados em tabelas e gráficos;
  - medidas descritivas (média, mediana, moda, quantis);
  - medidas de variação (desvio padrão, variância);
  - população vs. amostras;
  - distribuições, erro padrão e significância estatística;
  - introdução à programação estatística;
  - automatização da busca, organização e tratamento de dados empíricos; e
  - utilização tais dados para extrair estatísticas descritivas, fatos estilizados e análises gráficas.

Estrutura do curso

Como vamos aprender isso?

$curso = teoria + pratica$

- Teoria: teoria estatística, conceitos, compreensão acadêmica.
- Prática: programação aplicada à estatística.

Sumário

O que é estatística?

Como ela nos ajuda a compreender o mundo

- Estatística é um *conjunto de métodos* para planejar experimentos e obter dados e derivar conclusões de tal dados depois de:
  - organizá-los;
  - resumi-los;
  - analisá-los; e
  - interpretá-los.

Definições

Dados

- Dados são coleções de observações específicas sobre indivíduos, domicílio, países, máquinas, fábricas, etc.
- Exemplo. A *Pesquisa Nacional de Amostra de Domicílios*, realizada pelo IBGE anualmente, coleta uma série de informações sobre características de domicílios brasileiros:
  - Domicílio 1: {*UF*: DF; *Município*: Brasília; *Número de habitantes*: 3; *Número de cômodos*: 2; *Tem geladeira*: Sim; *Tem TV a Cores*: Sim; *Tem máquina de lavar*: Sim; *Renda familiar habitual no mês*: R$ 18.000,00; etc.};
  - Domicílio 2: {*UF*: MA; *Município*: São Luís; *Número de habitantes*: 5; *Número de cômodos*: 2; *Tem geladeira*: Sim; *Tem TV a Cores*: Sim; *Tem máquina de lavar*: Não; *Renda familiar habitual no mês*: R$ 2.000,00; etc.;
  - etc...

Definições

População, amostra e censo

- População: conjunto completo de indivíduos, objetos ou unidades sobre os quais queremos informações.
- Amostra: subconjunto de unidades da *população* acerca das quais realmente coletamos informações.
  - Importante: Um dos objetivos da ciência estatística é desenhar pesquisas estatísticas de modo a que as *amostras sejam representativas* - isto é, que as informações inferidas sobre a *amostra* representem as características da *população*.
- Censo: tentativa de amostrar toda a população.
  - Além do Censo Demográfico (que ocorre a cada 10 anos), há algum outro Censo no Brasil?

Definições

Variáveis, parâmetros e estatísticas

- Variáveis: características que medimos em cada unidade.
  - Exemplos: Salário; idade; e sexo de indivíduos. Cada uma dessas é uma *variável*.
- Parâmetros: características numéricas da população.
  - normalmente, são essas características que queremos saber, mas é muito difícil e custoso medir as características de toda a população ao longo de muitos períodos.
- Estatísticas: características numéricas da *amostra*.
  - Estatísticas (da amostra) são utilizadas para estimar os parâmetros (da população)

Definições

Tipos de variáveis

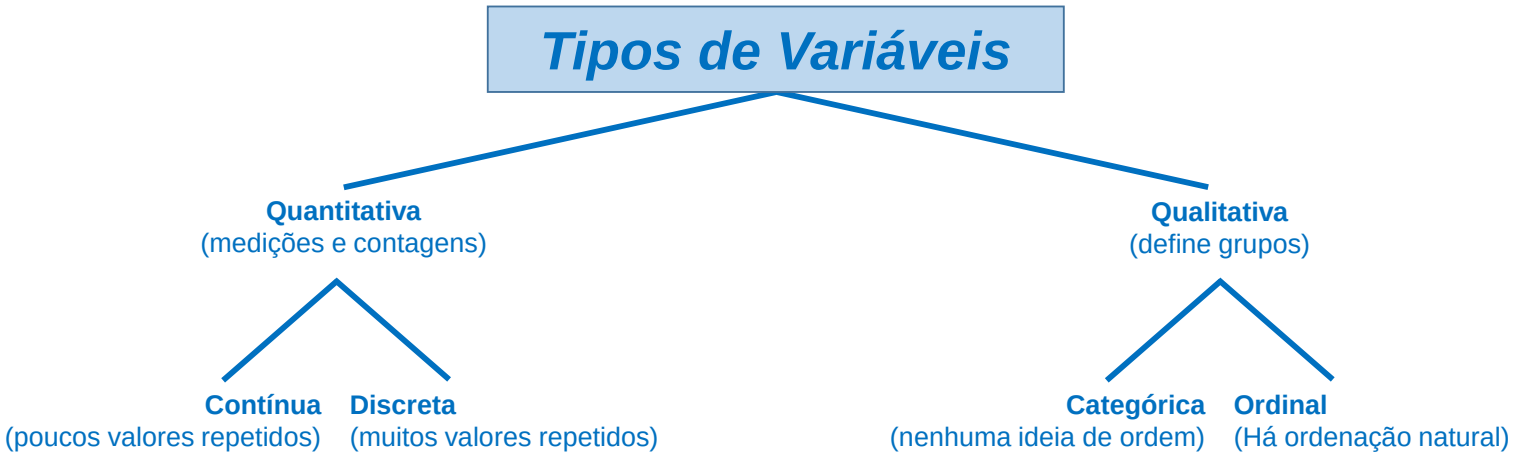

Figura: Tipos de variáveis. Detalhamento

Definições

Tipos de variáveis: quantitativas

- Variáveis quantitativas são aquelas que se referem a medições e contagens.
  - Variáveis discretas: são aquelas em que a contagem segue um intervalo pré-determinado. *Exemplo: quantos anos de estudo você completou?*
  - Variáveis contínuas: são aquelas em que a contagem inclui frações infinitesimais. *Exemplo: qual seu peso/altura?*.

Definições

Tipos de variáveis: qualitativas

- Variáveis qualitativas são aquelas que definem grupos que classificam os indivíduos da amostra.
  - Variáveis categóricas: são aquelas em que não há ideia de ordem entre as variáveis. *Exemplo: grupo sanguíneo; religião.*
  - Variáveis ordinais: são aquelas em que a contagem inclui frações infinitesimais. *Exemplo: classe - baixa/média/alta; escolaridade - fundamental/médio/superior*.

Problemas na análise estatística

Representatividade

- Para ser úteis, estatísticas têm de ser representativas.
  - Exemplo: se você pesquisar o que as pessoas acham do Lula (FHC) na frente do Congresso Nacional do PT (PSDB), o resultado da pesquisa não será representativo do conjunto da população.

Problemas na análise estatística

Vieses

- Viés de auto-seleção:
  - a REVISTA ÉPOCA faz uma enquete sobre *cobrança de imposto de igrejas* em seu *site*. O resultado é que 95% dos votantes é contrário. Logo depois, descobre-se que líderes religiosos pediram para fiéis votarem na enquete. Ela representa o conjunto da população brasileira?

Problemas na análise estatística

Vieses

- Viés de seleção:
  - fazem uma pesquisa que escolhe *aleatoriamente* (ou seja, sem viés de auto-seleção) pessoas para responder a seguinte pergunta: "A mudança climática é causada por fatores humanos?". 99% dos respondentes disseram que sim. Mas a pesquisa só inclui pesquisadores da NASA. Ela representa o conjunto da população americana?

Problemas na análise estatística

Vieses

- Viés causado pela formulação de pergunta perguntas:
  - Você concorda com a reforma trabalhista?
  - Você concorda com a *retirada de direitos promovida pela* reforma trabalhista?
  - Você concorda com a reforma trabalhista, *considerando que ela vai aumentar a probabilidade de trabalhadores pobres encontrarem um emprego formal*?

Problemas na análise estatística

A pesquisa estatística bem planejada

- Definir circunscrições de interesse
- Escolher unidades/indivíduos aleatoriamente dentro de cada área de interesse
- Definir pesos representativos com base em características demográficas conhecidas (região, religião, sexo, raça, etc...).

Problemas na análise estatística

A pesquisa estatística bem planejada

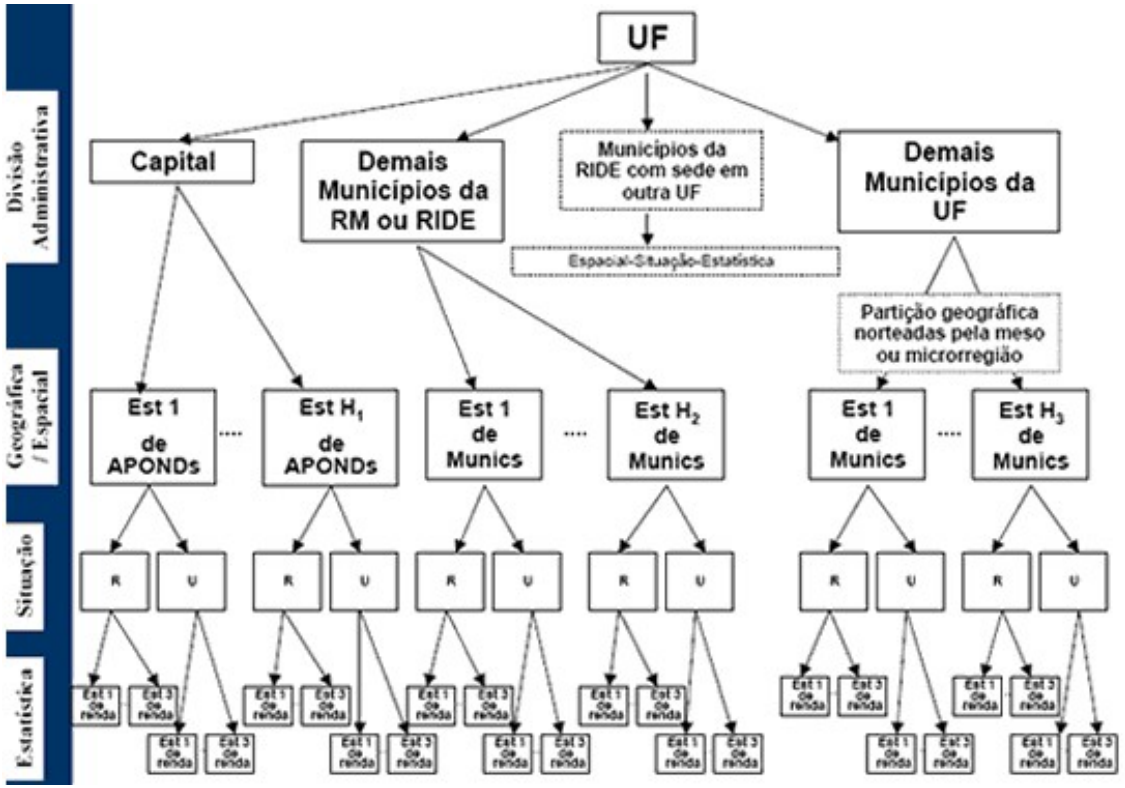

Figura: IBGE. Metodologia de estratificação da PNAD.

Problemas na análise estatística

O experimento estatístico ideal

- Escolher aleatoriamente unidades para teste
- Escolher aleatoriamente grupo de tratamento e grupo de controle
- Tratar um dos grupos e observar

Problemas na análise estatística

O experimento estatístico ideal

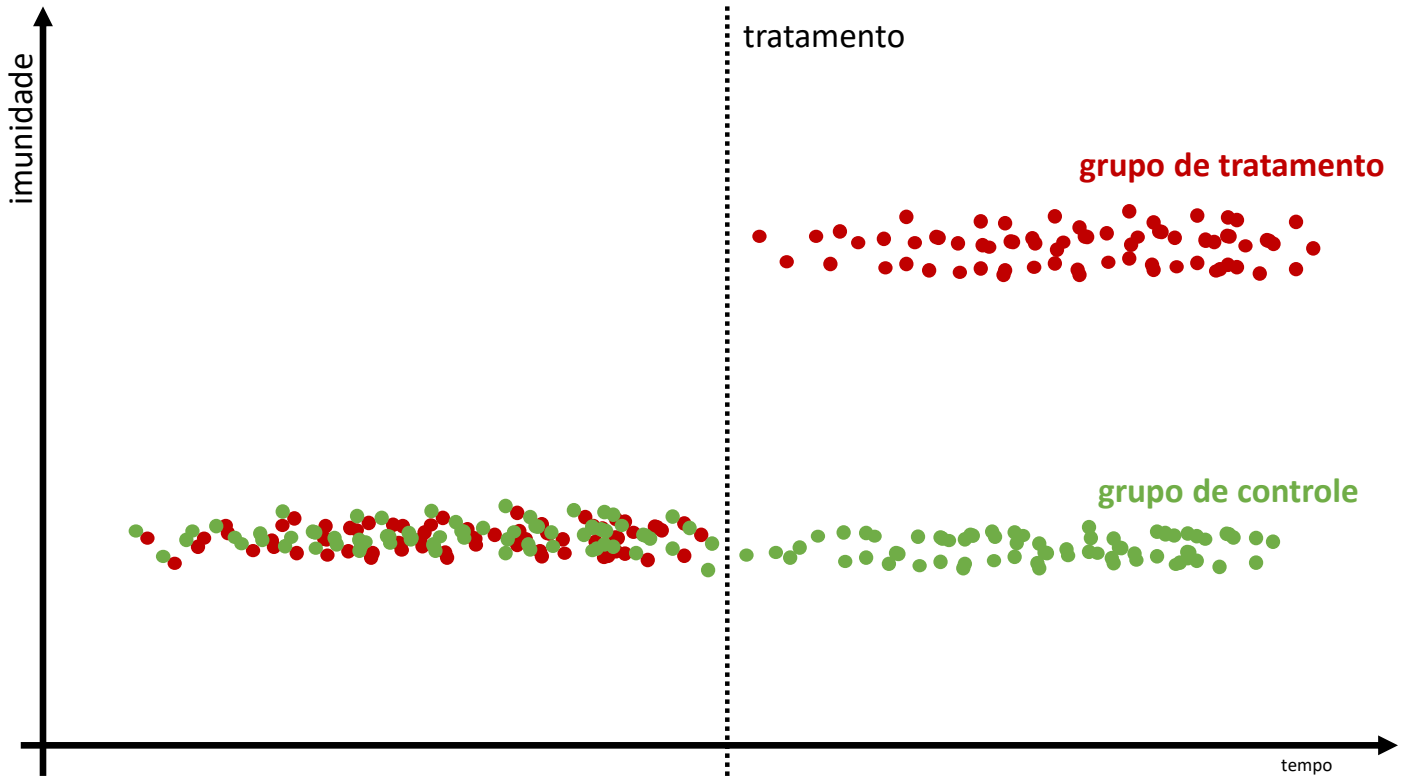

Figura: Experimento. Exemplo conceitual de experimento com grupos de tratamento e controle aleatorizados.

Problemas na análise estatística

E se não podemos fazer um experimento estatístico ideal?

- Nem sempre isso é possível
- Por isso, muitas vezes temos que fazer estudos "observacionais"
- E sempre lembrar que correlação não é causalidade

Problemas na análise estatística

Correlação não é causalidade

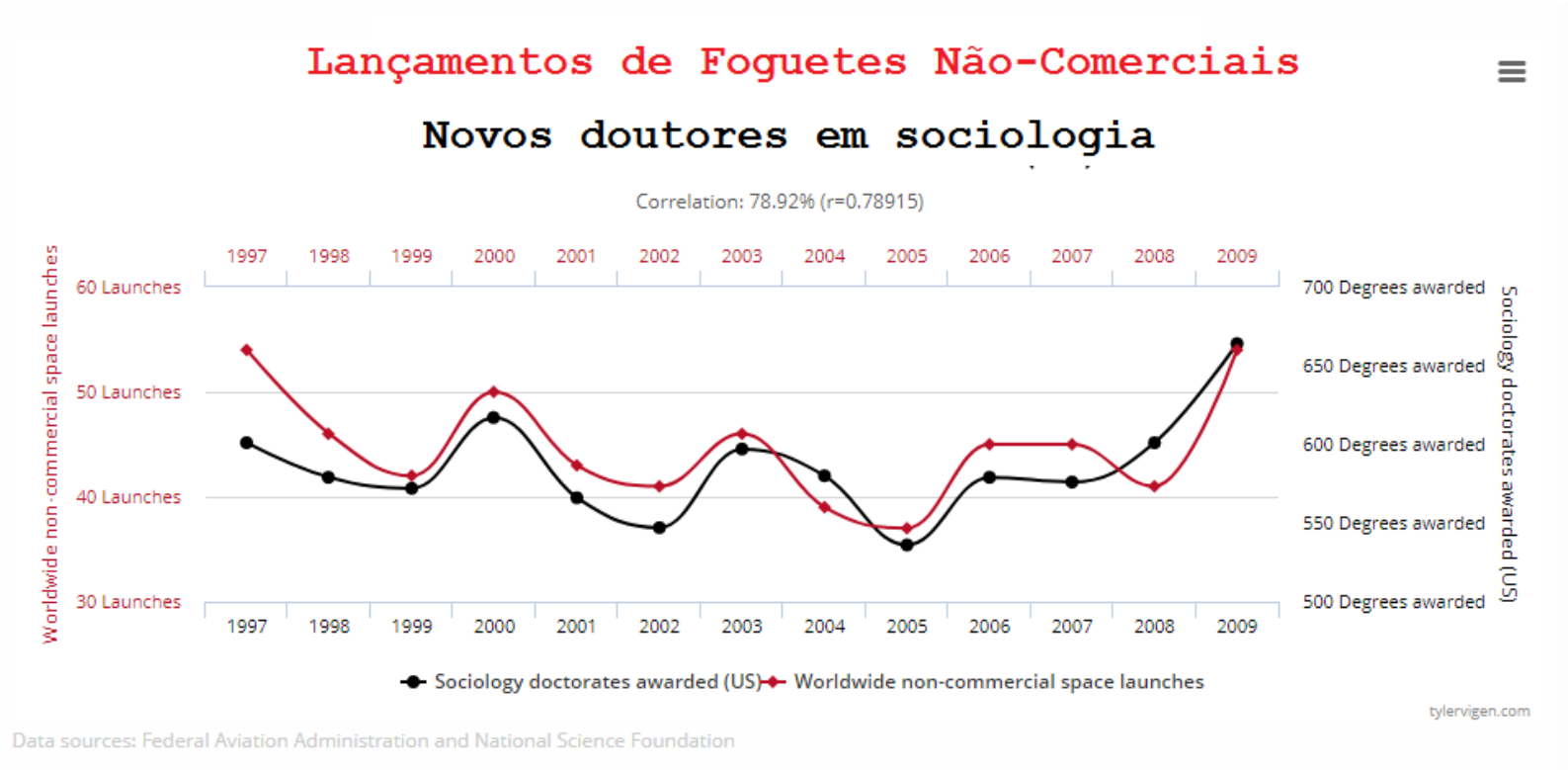

Figura: Correlação Espúria. Exemplo de como correlação nem sempre é indício causalidade.

Problemas na análise estatística

E se não podemos fazer um experimento estatístico ideal?

- Uma alternativa é buscar métodos "experimentos naturais".
- Exemplos:
  - Refugiados com características similares são alocados aleatoriamente em cidades diferentes
  - Em uma cidade que está na divisa de dois estados diferentes, a lei de um dos estados aumenta o salário mínimo em parte da cidade mas não no outro
  - Por causa de um "acidente histórico", um país se torna metade capitalista e metade socialista

Problemas na análise estatística

Experimentos naturais

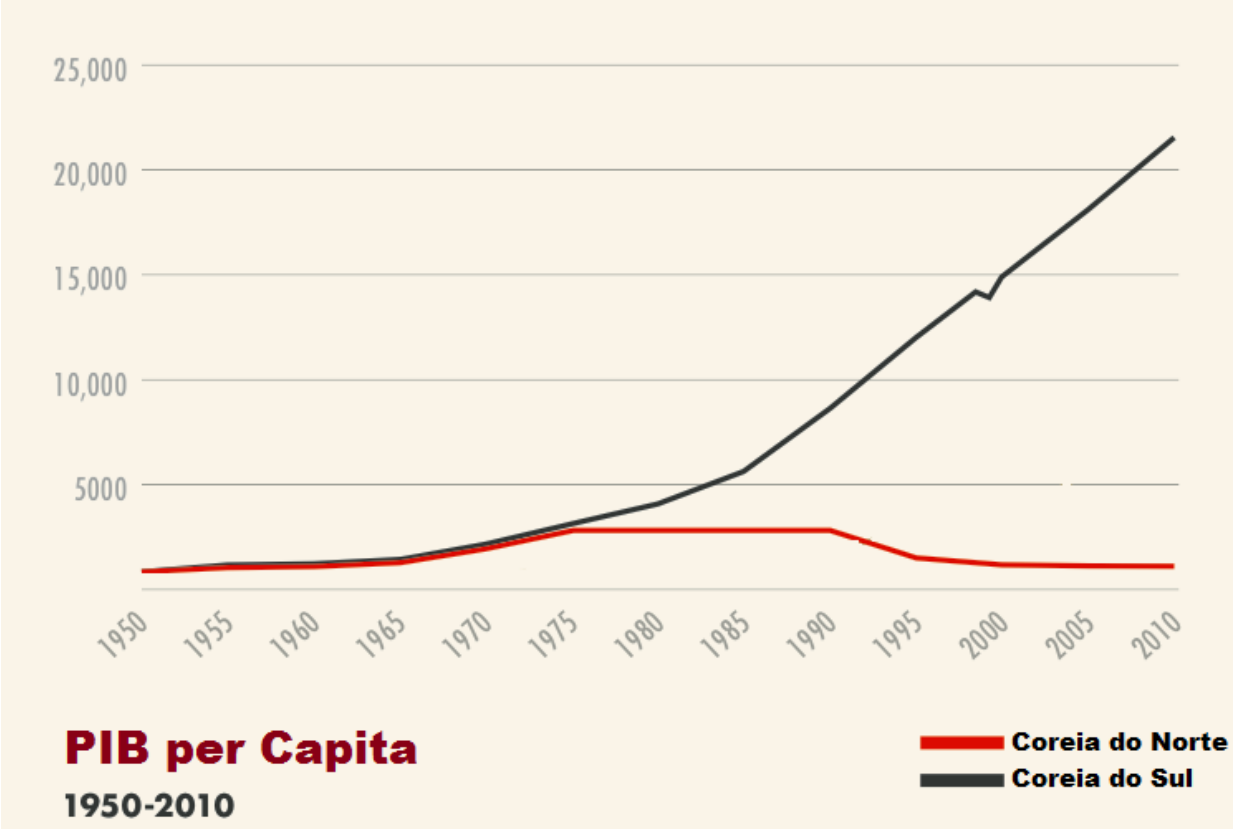

Figura: Experimentos Naturais. A divisão da Coreia em dois países e sua diferente performance econômica é, muitas vezes, tido como um experimento natural de modelos econômicos.

Sumário

Por que Python?

Python para ciência de dados

- Python é uma linguagem de programação geral que tem crescente popularidade tanto em ciência de dados quanto para outros propósitos.
- Python prioria a interpretação do código por seres humanos - e é muito fácil para quem está aprendendo.
- Segue princípios como (ver *Zen of Python*):
  - O explícito é melhor do que o implícito;
  - O simples é melhor do que o complexo;
  - Facilidade de leitura importa.

Por que o Python?

Fácil compreensão

Java

```
class myprog
{
    public static void main(String args[])
    {
        System.out.println("Oi!");
    }
}
```

Python

```
print("Oi!")
```

Por que o Python?

R vs. Python

Figura: R vs. Python. Qual é a melhor linguagem?

Por que o Python?

Comparação entre linguagens

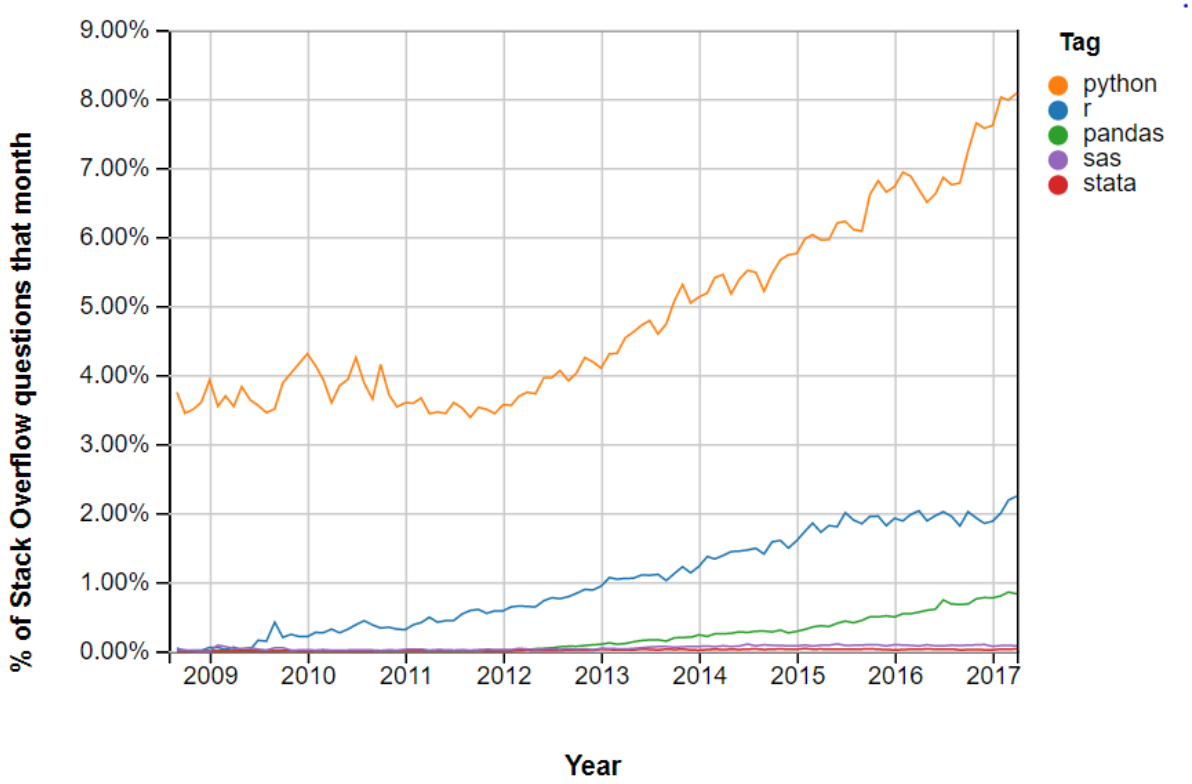

Figura: Linguagens de ciência de dados. Tendências de popularidade

Introdução ao Spyder

O que é o Spyder?

- Spyder (*Scientific PYthon Development EnviRonment*) é um ambiente de desenvolvimento interativo que facilita a programação de programas específicos para ciência de dados.
- Como ele facilita a nossa vida?
  - Avisa quando seu código está errado;
  - Facilita a compreensão de seu código por meio de coloração;
  - Mostra o seu código (editor) e o resultado (console de comando) dele na mesma janela;
  - Exibe quais objetos estão gravadas na memória, após a você rodar seu código parcialmente;
  - Tem ferramentas fáceis para acessar a ajuda/documentação, quando você não sabe direito o que fazer.

Introdução ao Spyder

Como usar o Spyder?

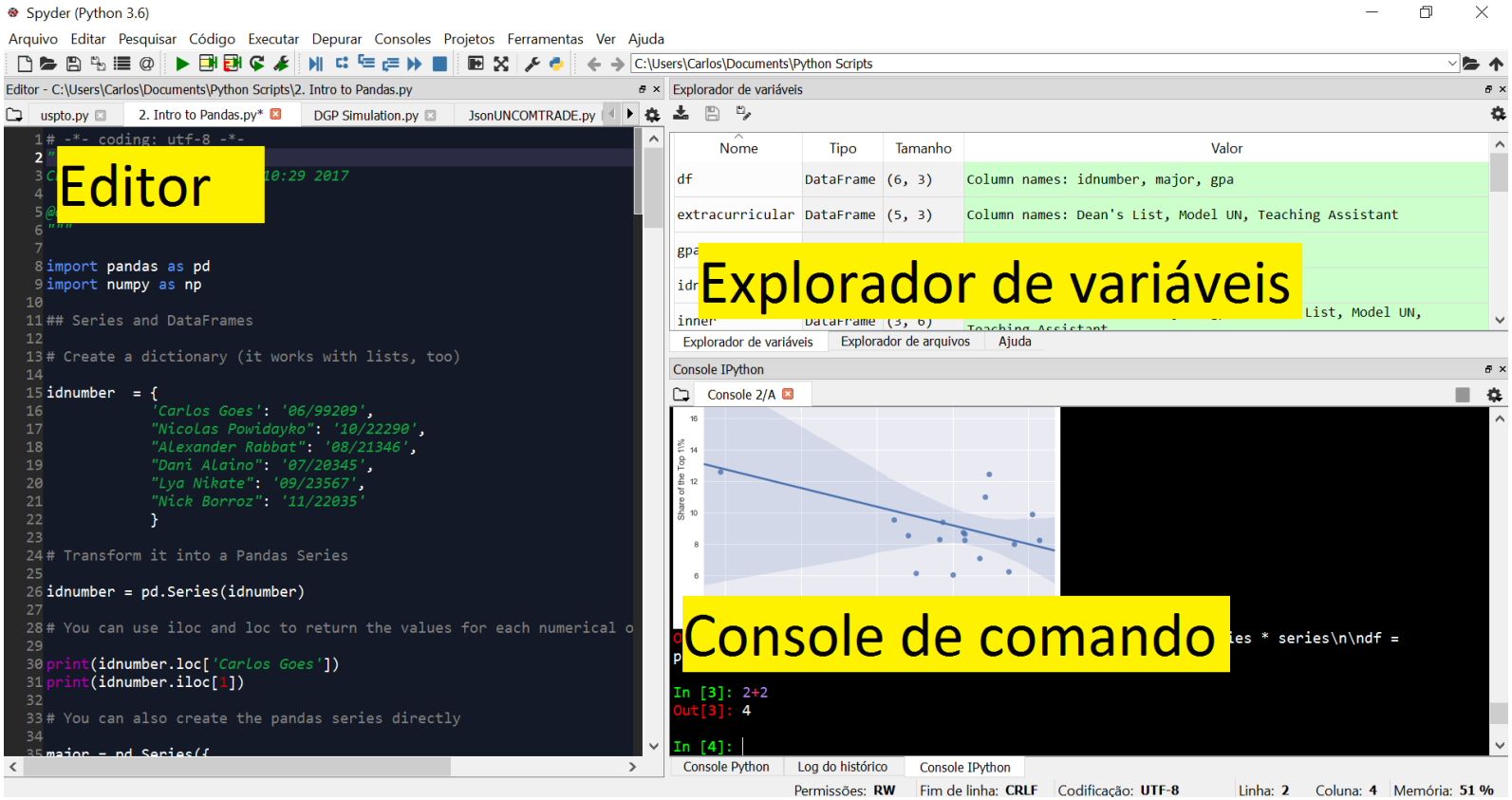

Figura: Spyder. Componentes do IDE.

Integers, floats, strings e booleans

Integers

- `Integers` são números *inteiros* (sem decimais ou frações).

$$\mathbb{Z} \equiv \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\} \quad (1)$$

- Exemplo no Python:

```
var = 20
print(type(var))
```

Integers, floats, strings e booleans

Floats

- `Floats` são representações aritméticas de todos os números *reais*

$$\begin{aligned} \mathbb{R} &\equiv \{\ldots, -\pi, -2.5, -\sqrt{2}, -\frac{1}{2}, 0, \frac{1}{2}, \sqrt{2}, 2.5, \pi, \ldots\} \quad (2) \\ \mathbb{R} &\equiv \{\ldots, -3.14159265, -2.5, -1.41421356, \\ &\quad -0.5, 0, 0.5, 1.41421356, 2.5, 3.14159265, \ldots\} \end{aligned}$$

- Exemplo no Python:

```
var = 3.14159265
print(type(var))
```

Integers, floats, strings e booleans

Strings

- `Strings` são representações de texto.
- Exemplo no Python:

```
var = "Esse é um string"
print(type(var))
```

Integers, floats, strings e booleans

Booleans

- `Booleans` são representações lógicas (verdadeiro ou falso).
- Exemplo no Python:

```
var1 = True
print(type(var1))
var2 = (2+2 == 5)
print(type(var2))
```

Integers, floats, strings e booleans

Tipos de variáveis e categorizações teóricas

- Variáveis quantitativas discretas: `integers`.
- Variáveis quantitativas contínuas: `floats`.
- Variáveis qualitativas categóricas: `booleans` ou `strings`.
- Variáveis qualitativas ordinais: `strings` ou `integers`.

Integers, floats, strings e booleans

Operações numéricas

- Adição:

```
x + y
```

- Subtração:

```
x - y
```

- Divisão:

```
x / y
```

- Multiplicação:

```
x * y
```

- Exponenciação:

```
x ** y
```

- Resto:

```
x % y
3 % 2 is 1
```

Integers, floats, strings e booleans

Operações com palavras

- Adição:

```
str1 = "Carlos"
str2 = "Góes"
print(str1 + " " + str2)
```

- Multiplicação:

```
str1 = "a"
print(str1 * 10)
```

Integers, floats, strings e booleans

Operações lógicas

- Igualdade:

```
2 + 2 is 4
2 + 2 == 4
str = "a"
str is "a"
str == "a"
```

- Desigualdade:

```
2 + 2 is not 4
2 + 2 != 4
str = "a"
str is not "a"
str != "a"
```

- Maior ou menor que:

```
10 > 100
10 < 100
```

Integers, floats, strings e booleans

Operações lógicas

- Em operações lógicas, `True` tem o valor de 1 e `False` tem o valor de 0. Assim, é possível utilizar esse conceito para realizar operações matemáticas.
- Soma (quantos são verdadeiros?):

```
a = 1 + 1 is 2
b = 2 * 2 is 4
c = 2 * 2 is 5
d = a + b + c
print(d)
```

- Multiplicação (todos são verdadeiros?):

```
a = 1 + 1 is 2
b = 2 * 2 is 4
c = 2 * 2 is 5
d = a * b * c
print(d)
```

Listas e dicionários

Listas

- `Lists` são conjuntos de variáveis ordenadas. Elas podem conter `integers`, `floats`, `strings` ou `booleans`. Listas são definidas sempre entre `[colchetes]` e seus elementos são separados por virgulas:

```
lista1 = [2, 20.5, "Oi!", 10 < 100]
print(lista1)
```

- `Lists` são indexadas de 0 ao elemento $n-1$ da lista, sendo que n é o número de elementos. Você pode acessar um elemento com a sintaxe `nome_da_lista[indexador]`:
- Por exemplo, `print(lista1[0])` retorna `2`.
- Equanto, `print(lista1[3])` retorna `True`.

Listas e dicionários

Listas

- Você pode alterar um elemento de uma lista atribuindo um valor a um indexador específico:

```
lista1[0]= 20
print(lista1)
```

- Adicionar um elemento à lista:

```
lista1.append(79.2)
print(lista1)
```

- Usar multiplicação para repetir o conteúdo da lista:

```
lista2 = lista1 * 2
print(lista2)
```

- Ou usar adição para juntar duas listas:

```
lista3 = [1,2]
lista4 = [5,6]
lista5 = lista3 + lista4
```

Listas e dicionários

Dicionários

- `Dictionaries` são, como `lists`, conjuntos de variáveis. Mas, de forma distinta, as informações dentro de dicionários não são indexadas por chaves (normalmente palavras):

```
pessoa1 = {
          'nome': 'Milton Friedman',
          'nascimento': '31/07/1912'
          }
print(pessoa1)
print(pessoa1['nome'])
```

- Adicionar um elemento ao dicionário:

```
pessoa1.update({'nacionalidade': 'EUA'})
print(pessoa1)
```