Ciência de Dados

Amiraldo Ferreira, Jheickson Felipe e Lorenzo Alberto

Santarém, Pará
17 Janeiro 2023

LibreOffice

Agenda

- O que é Ciência de Dados?
- Linguagens para Ciências de Dados.
- R e RStudio, Jupyter R/Python.

Illustrations by Pixeltrue on icons8

O que é Ciência de Dados?

- Extração de dados significativos;
- Combinação de princípios matemáticos e estatísticos com suporte da inteligência artificial e engenharia da computação;
- Análise de grandes quantidades de dados.

Qual a importância?

- Armazenamento automático de informações;
- Reduz a redundância e impulsiona a inovação;
- Produz resultados completos, como:
  - Códigos,
  - Resultados numéricos; e,
  - Relatórios.

O que faz um cientista de dados?

- Previsão do futuro com base matemática;
- Desenvolve estratégias para analisar dados;
- Prepara, explora e visualiza dados que sejam úteis a um determinado contexto.

Análises da Ciência de Dados

Processos da Ciência de Dados

Técnicas mais utilizadas

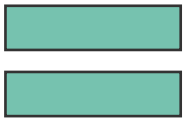
- Classificação;
- Regressão; e,
- Clustering.

Uso em tecnologias

- Inteligência Artificial;
- Computação em nuvem;
- Internet das Coisas; e,
- Computação quânticas.

Linguagens para Ciência de Dados

Ferramentas do Cientista de Dados



julia

Scala

SQL

GO

Swift

As mais populares

Python

- Altamente popular
- Código aberto
- Grande quantidade de Libraries

Python

- Altamente popular
- Código aberto
- Grande quantidade de Libraries

- Funções matemáticas avançadas

- Data Manipulation a partir de banco de dados

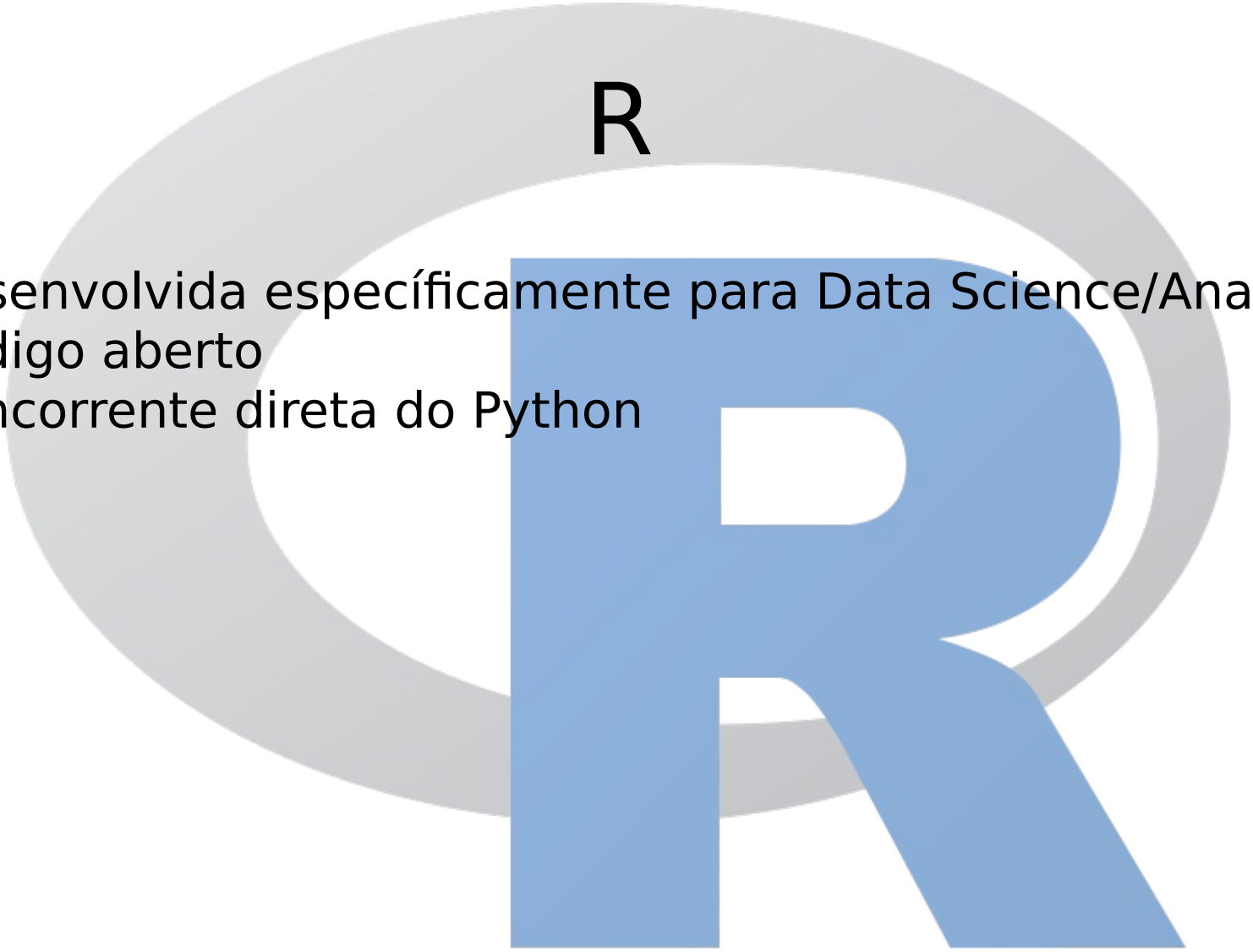
- Gráficos e visualização interativa

- Machine/Deep Learning

- Neural Networks

R

- Desenvolvida específicamente para Data Science/Analysis
- Código aberto
- Concorrente direta do Python



R

- Desenvolvida específicamente para Data Science/Analysis
- Código aberto
- Concorrente direta do Python

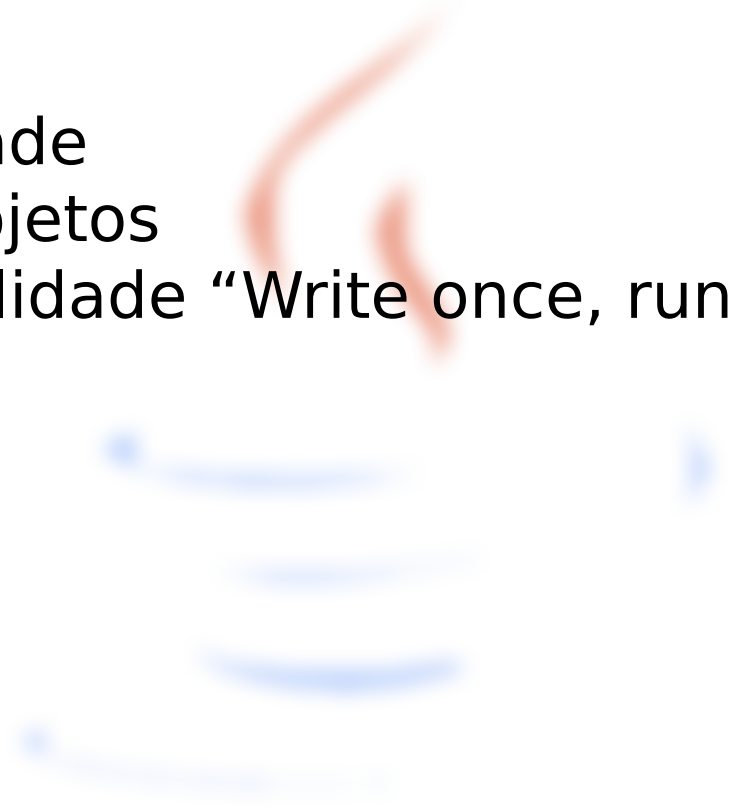
SQL

- Mais que uma linguagem de criação de bancos de dados relacionais
- Ganhando atenção na área de Ciência de Dados
- Scripts para atualização de dados

- Criação de relatórios
- Criação de gráficos
- Criação de painéis

Java

- Alta versatilidade
- Orientada a objetos
- Alta compatibilidade “Write once, run anywhere”



Java

- Alta versatilidade
- Orientada a objetos
- Alta compatibilidade “Write once, run anywhere”

- Uso de Pipelines e Parallel Processing

- Visualização de dados

Java

- Alta versatilidade
- Orientada a objetos
- Alta compatibilidade “Write once, run anywhere”

- Uso de Pipelines e Parallel Processing

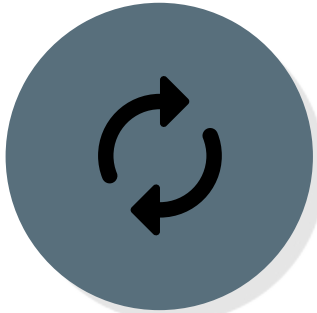
- Visualização de dados

Menções

Pondo a Mão na Massa

R

Linguagem de manipulação, análise e visualização de dados.



Jupyter Notebook

Ambiente de anotações, códigos e análise de dados.

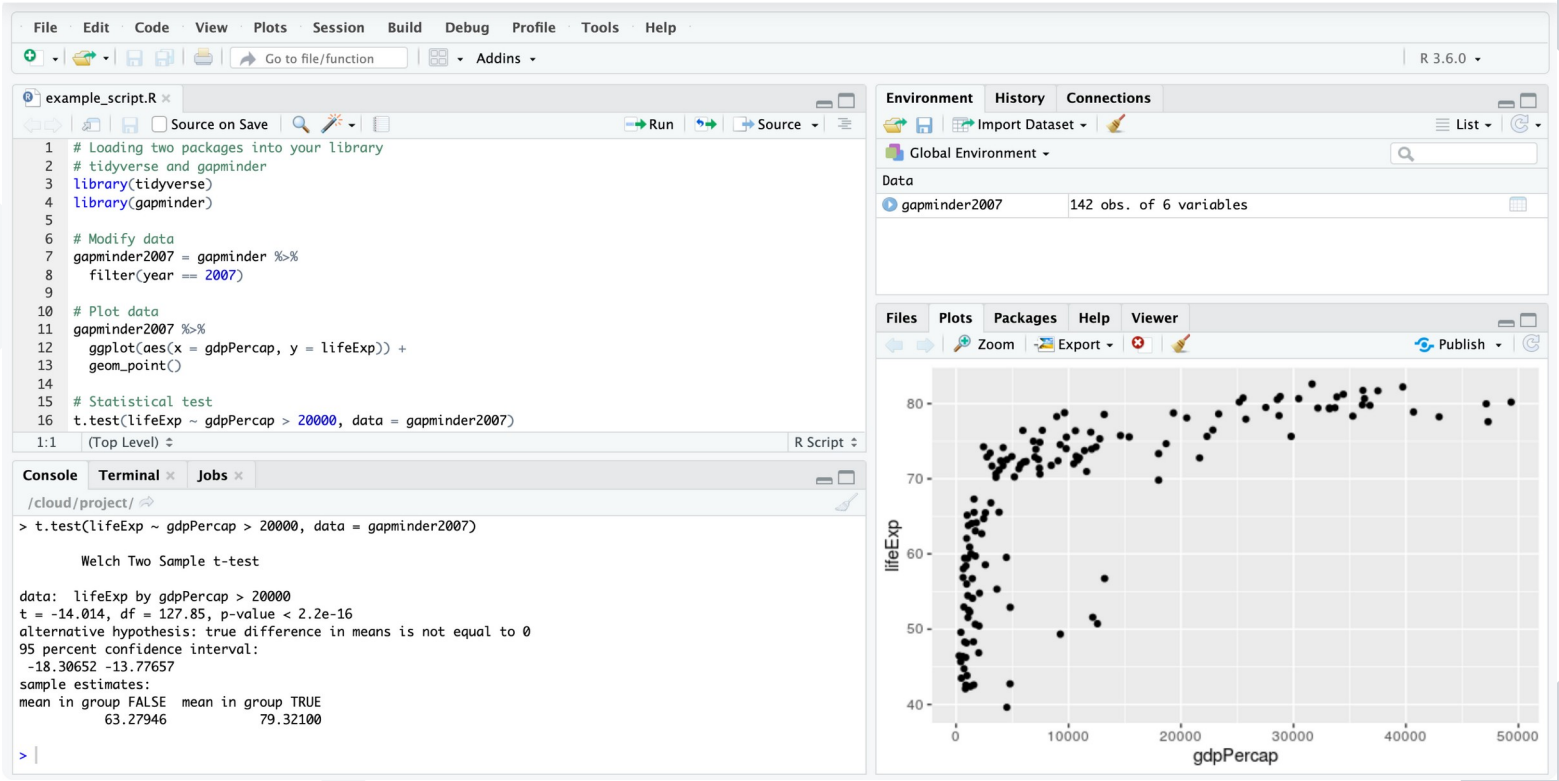
Python

Linguagem de programação de alto nível e de uso geral – fácil de entender e escrever.

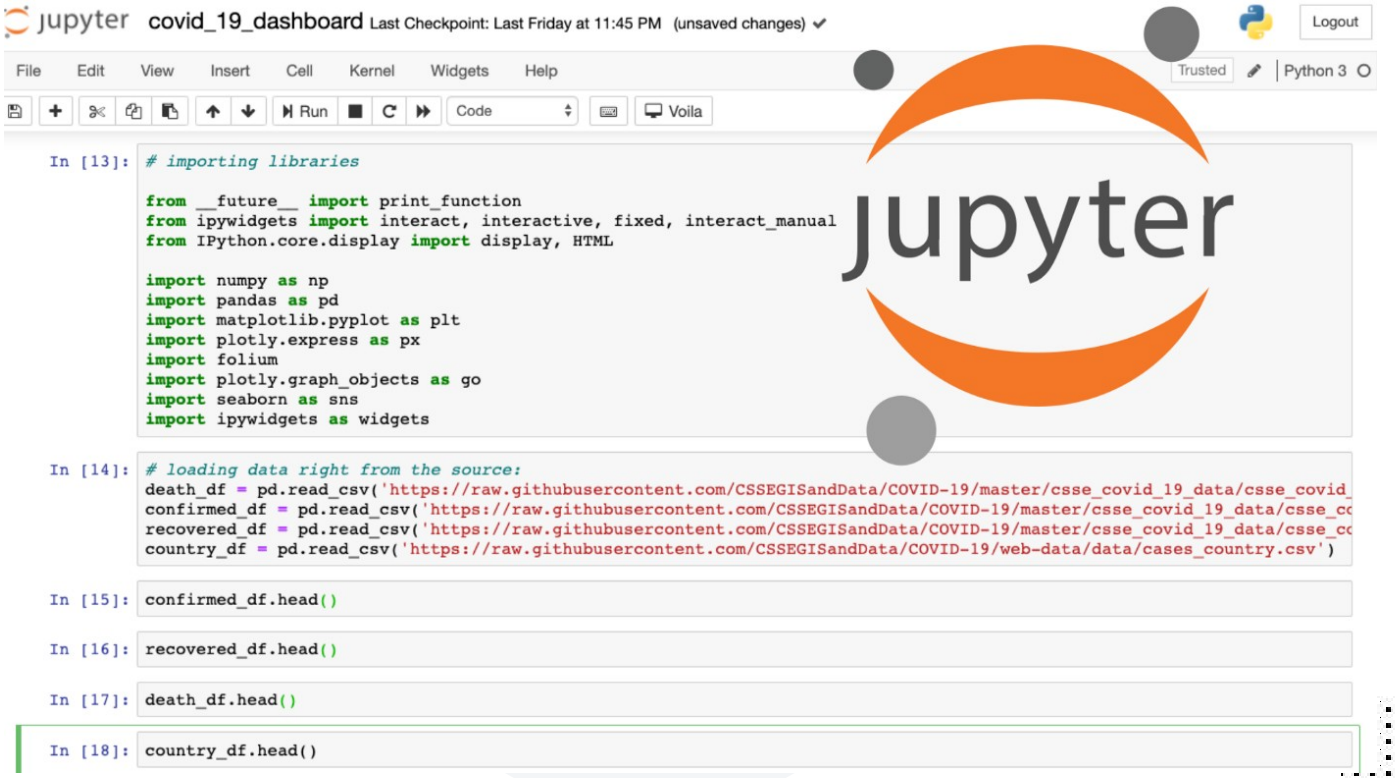
R – Ross & Robert

- Linguagem multi-paradigma orientada a objetos, programação funcional, dinâmica, fracamente tipada;
- Voltada à manipulação, análise e visualização de dados;
- Criada no departamento de Estatística da Universidade de Auckland, Nova Zelândia;
- Pesquisas mostram que a popularidade do R aumentou nos últimos anos.

R e RStudio


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.6.0
example_script.R
Source on Save
Run
Source
1 # Loading two packages into your library
2 # tidyverse and gapminder
3 library(tidyverse)
4 library(gapminder)
5
6 # Modify data
7 gapminder2007 = gapminder %>%
8 filter(year == 2007)
9
10 # Plot data
11 gapminder2007 %>%
12 ggplot(aes(x = gdpPercap, y = lifeExp)) +
13 geom_point()
14
15 # Statistical test
16 t.test(lifeExp ~ gdpPercap > 20000, data = gapminder2007)
1:1 (Top Level) R Script
Console Terminal Jobs
/cloud/project/
> t.test(lifeExp ~ gdpPercap > 20000, data = gapminder2007)
Welch Two Sample t-test
data: lifeExp by gdpPercap > 20000
t = -14.014, df = 127.85, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-18.30652 -13.77657
sample estimates:
mean in group FALSE mean in group TRUE
63.27946 79.32100
>
Environment History Connections
Import Dataset
List
Global Environment
Data
gapminder2007 142 obs. of 6 variables
Files Plots Packages Help Viewer
Zoom
Export
Publish
lifeExp
80
70
60
50
40
0
10000
20000
30000
40000
50000
gdpPercap

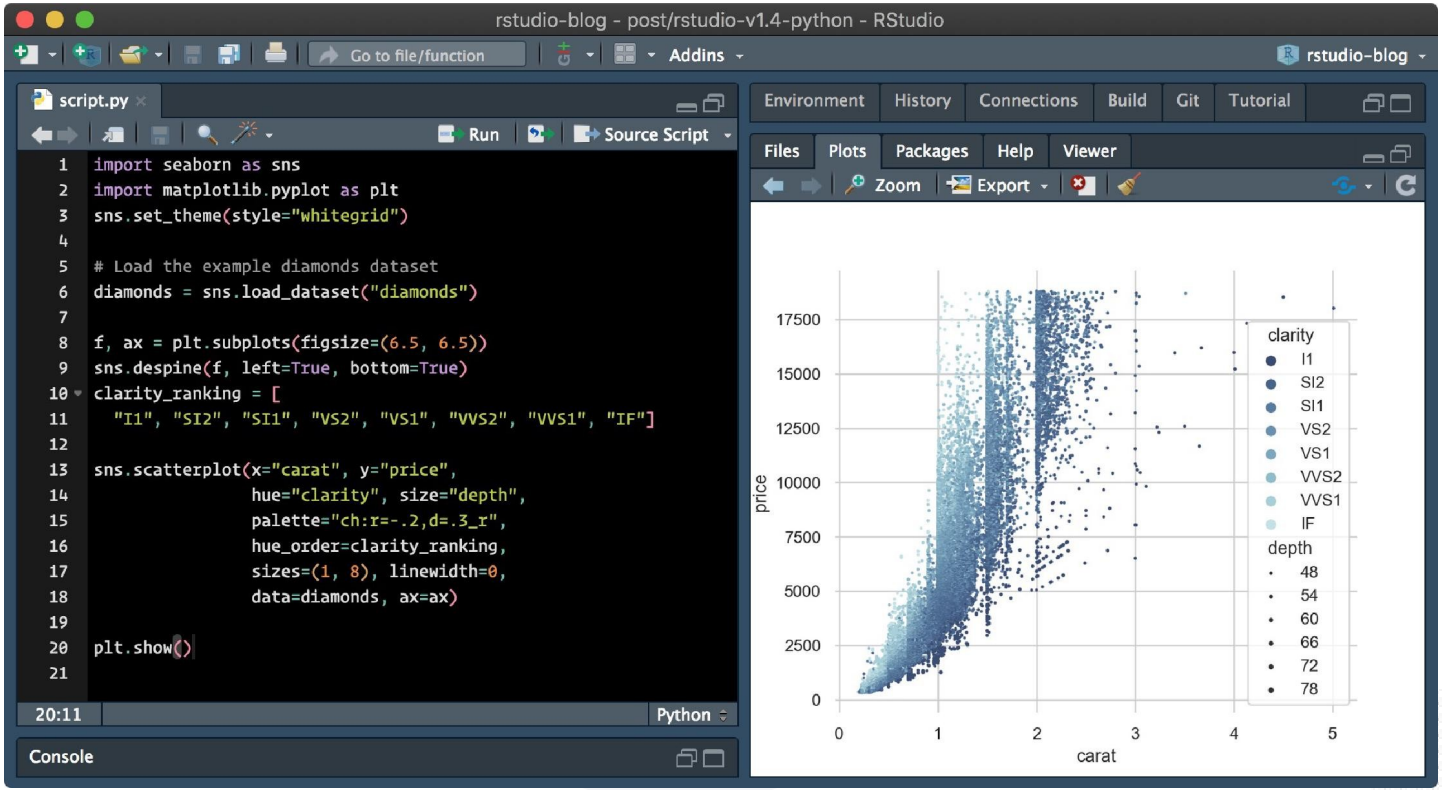
Jupyter Notebook



Python

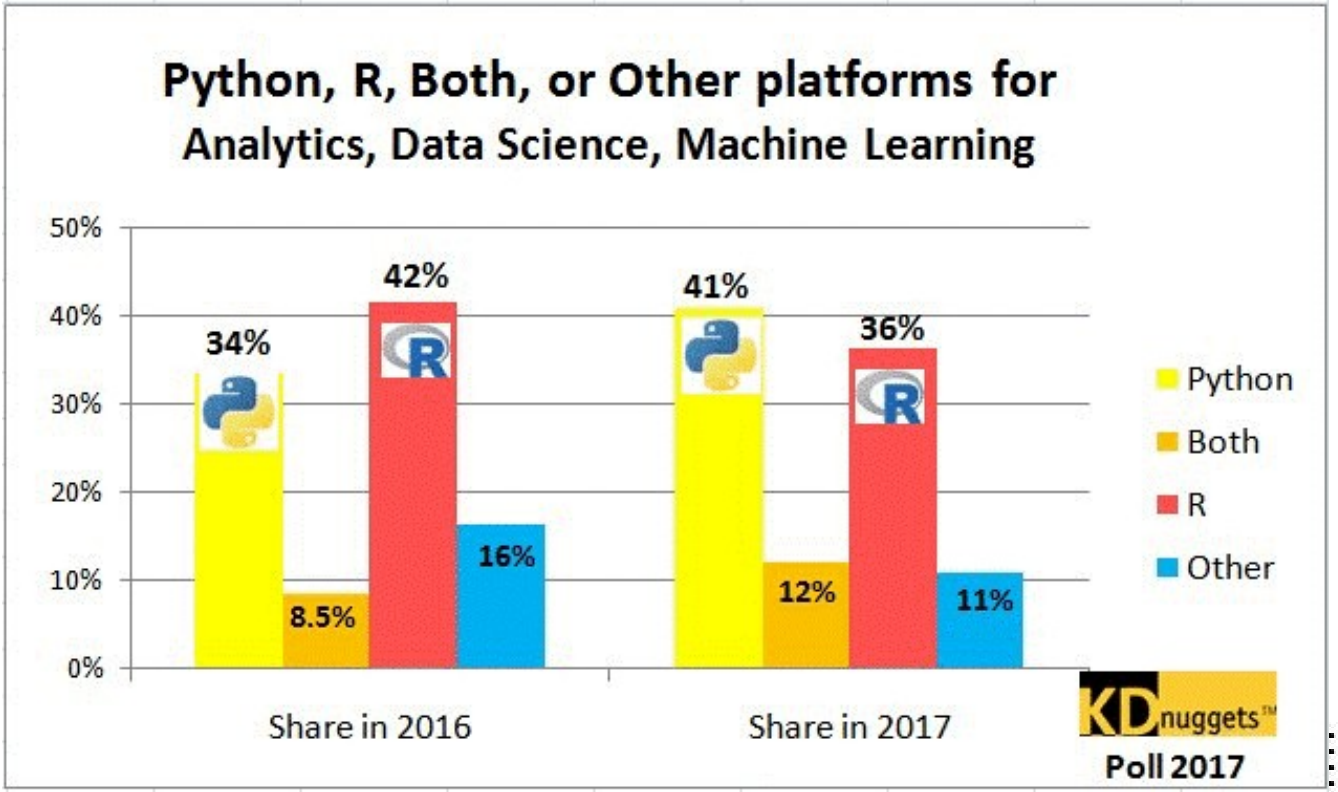
- Python is a high-level, general-purpose programming language. Designed for code readability with the use of significant indentation;
- Dynamically-typed and garbage-collected. Supports multiple programming paradigms;
- Comprehensive standard library;
- One of the most popular programming languages;

Python e Código

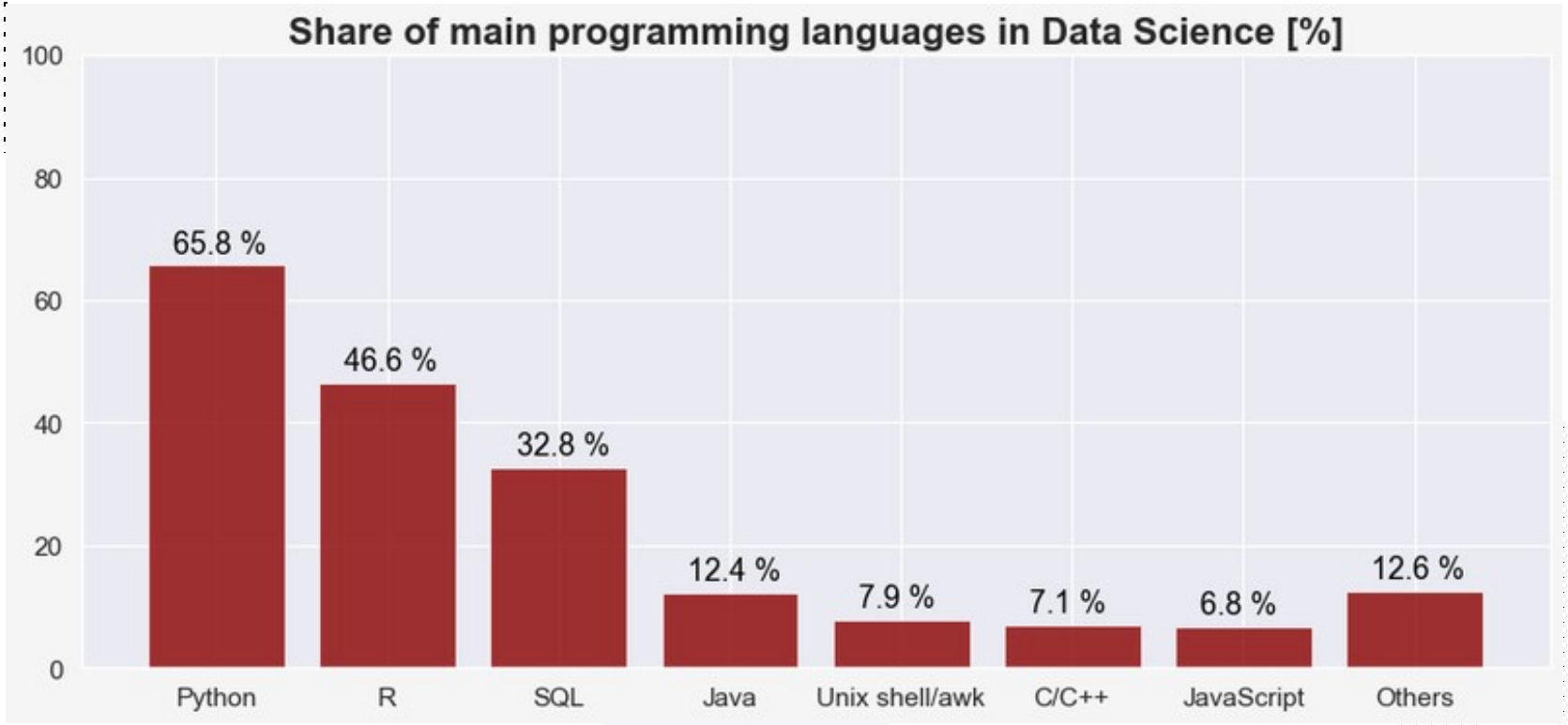


R e Python em Ciência de Dados

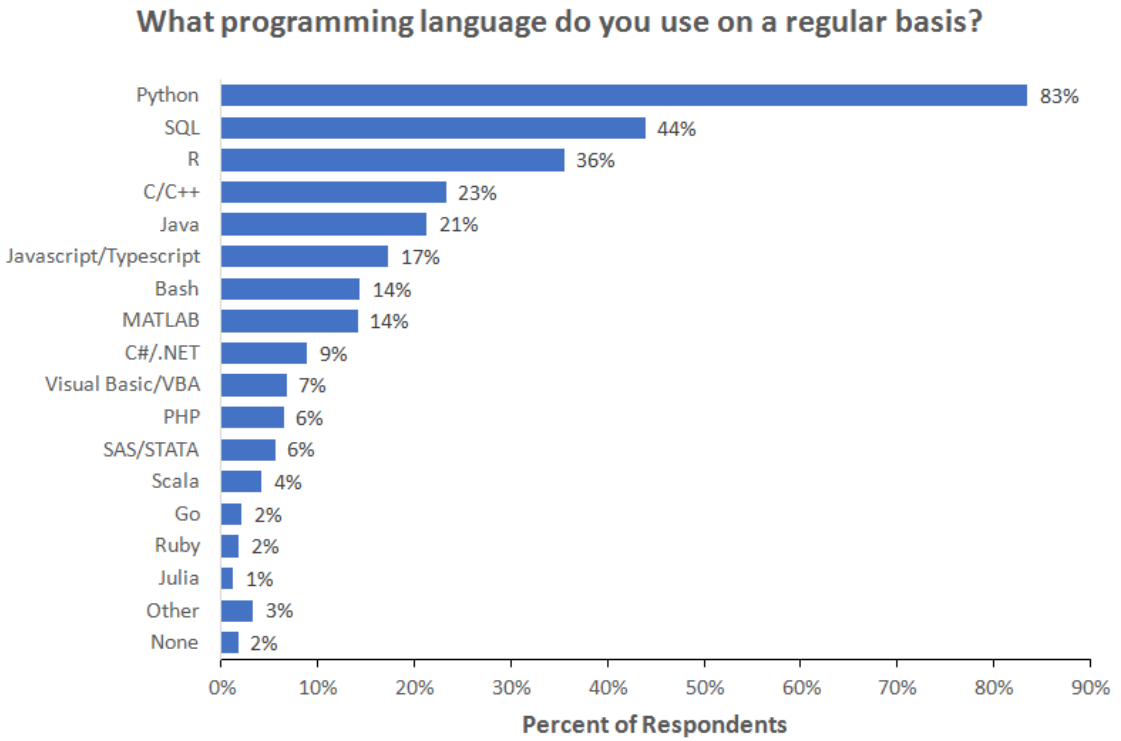
Cenário 2016/2017



Cenário 2019



Reflexo no Kaggle



Então é Python?

Não…..talvez.

Comparações gerais

Quesito	R	python
Facilidade de aprendizado	9,5/10,0 ✓	9,0/10,0
Manipulação de dados	9,5/10,0 ✓	9,0/10,0
Machine Learning	8,5/10,0	9,5/10,0 ✓
Visualização de dados	9,0/10,0 ✓	9,0/10,0 ✓
IDE	9,5/10,0 ✓	8,5/10,0
Comunidade	8,5/10,0	9,5/10,0 ✓
Velocidade	8,5/10,0	9,0/10,0 ✓
Empregos	6,5/10,0	9,5/10,0 ✓
Utilização comercial	5,5/10,0	9,5/10,0 ✓
Nuvem	9,0/10,0	9,5/10,0 ✓

Conclusão

average R user:

average Python enjoyer:

Por que não ambos?

A melhor linguagem é aquela que resolve seu problema.

Recomendações

- Google Colab;
- PyCharm;
- Spyder.