

# **Ciência de Dados**

Amiraldo Ferreira, Jheickson  
Felipe e Lorenzo Alberto

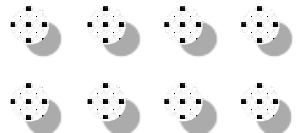
Santarém, Pará  
17 Janeiro 2023

# 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.

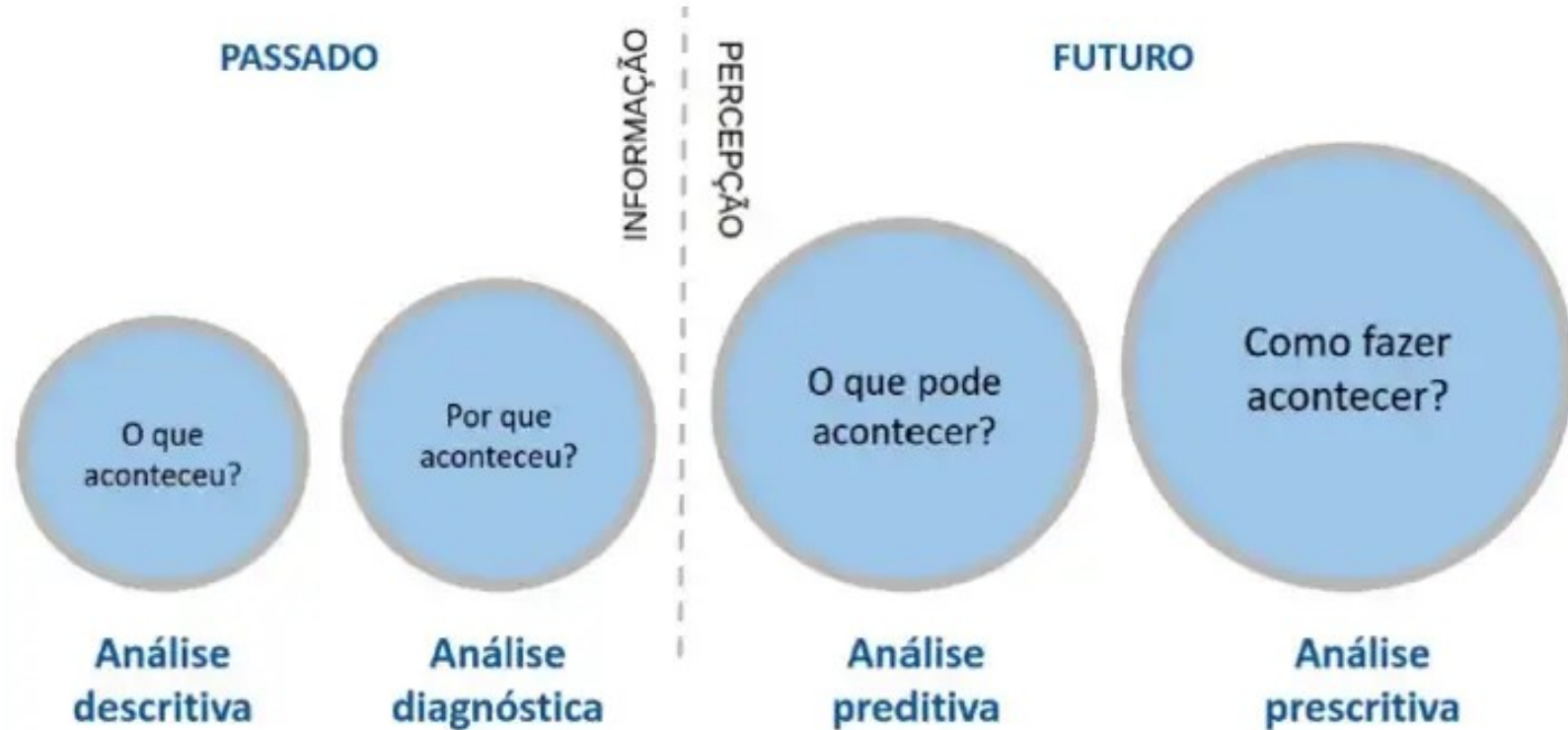
# Qual a importância?

- 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

**O**



**S**



**E**



**M**



**N**



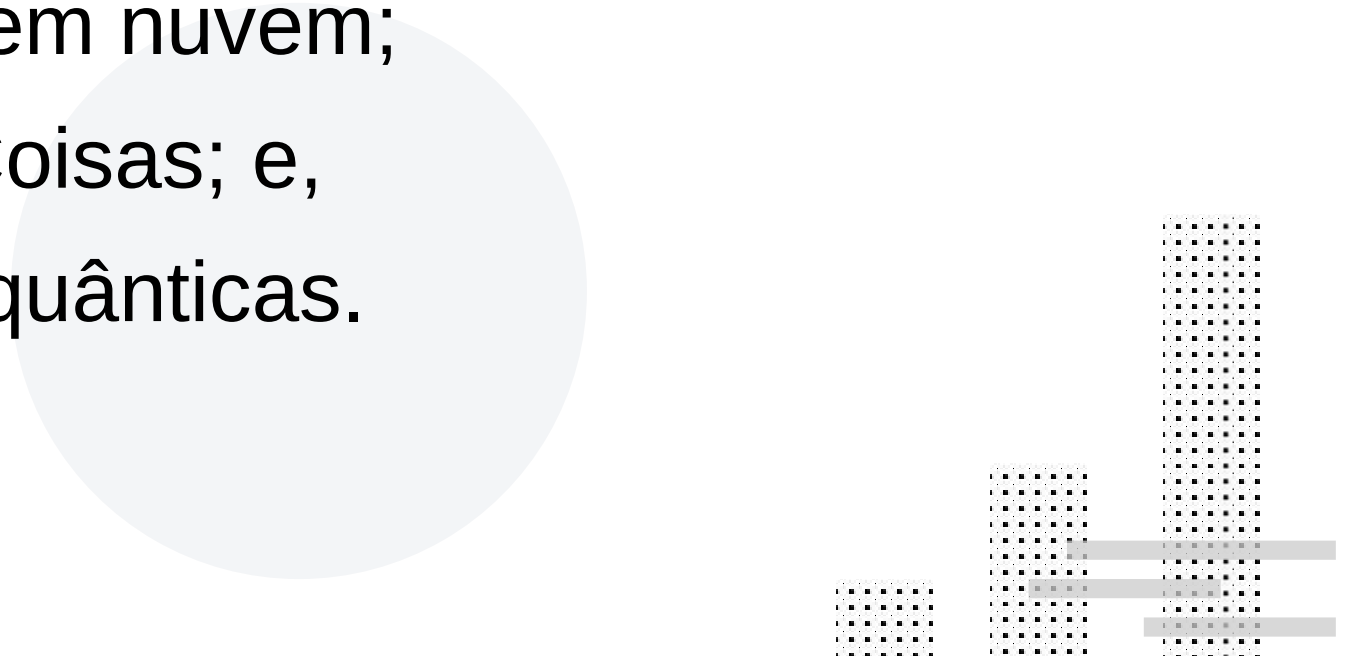
# Técnicas mais utilizadas

- Classificação;
- Regressão; e,
- Clustering.



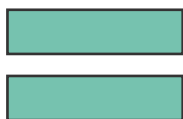


# Uso em tecnologias

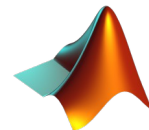
- Inteligência Artificial;
  - Computação em nuvem;
  - Internet das Coisas; e,
  - Computação quânticas.
- 

# Linguagens para Ciência de Dados

Ferramentas do Cientista de Dados



julia



Scala



SQL

GO



Swift

# As mais populares



# Python

The Python logo, consisting of two interlocking snakes, one blue and one yellow, is positioned in the background of the slide.

- 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

A large, light gray watermark of the R logo is centered in the background. It consists of a large capital 'R' with a smaller capital 'R' nested inside its upper loop.

# R

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

# R

- Desenvolvida especificamente 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

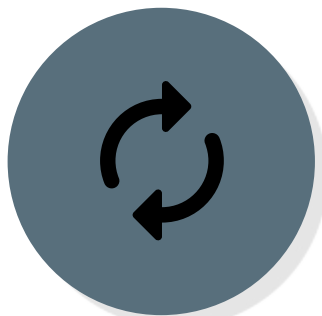


# Pondo a Mão na Massa



**R**

Linguagem de manipulação, análise e visualização de dados.



**Jupyter Notebook**

Ambiente de anotações, códigos e análise de dados.



**Python**

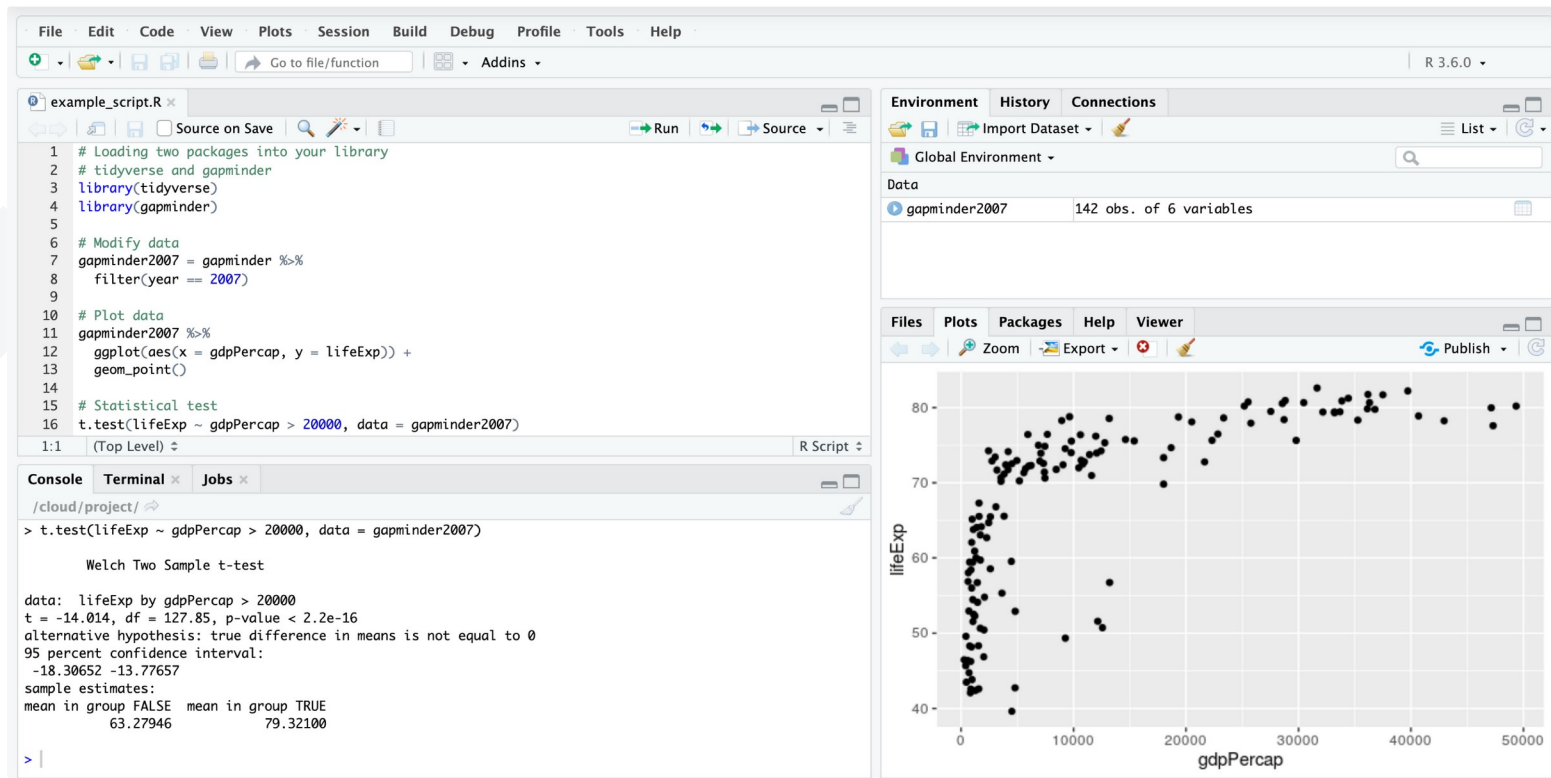
Linguagem de programação de alto nível e de uso geral – fácil de entender e escrever.



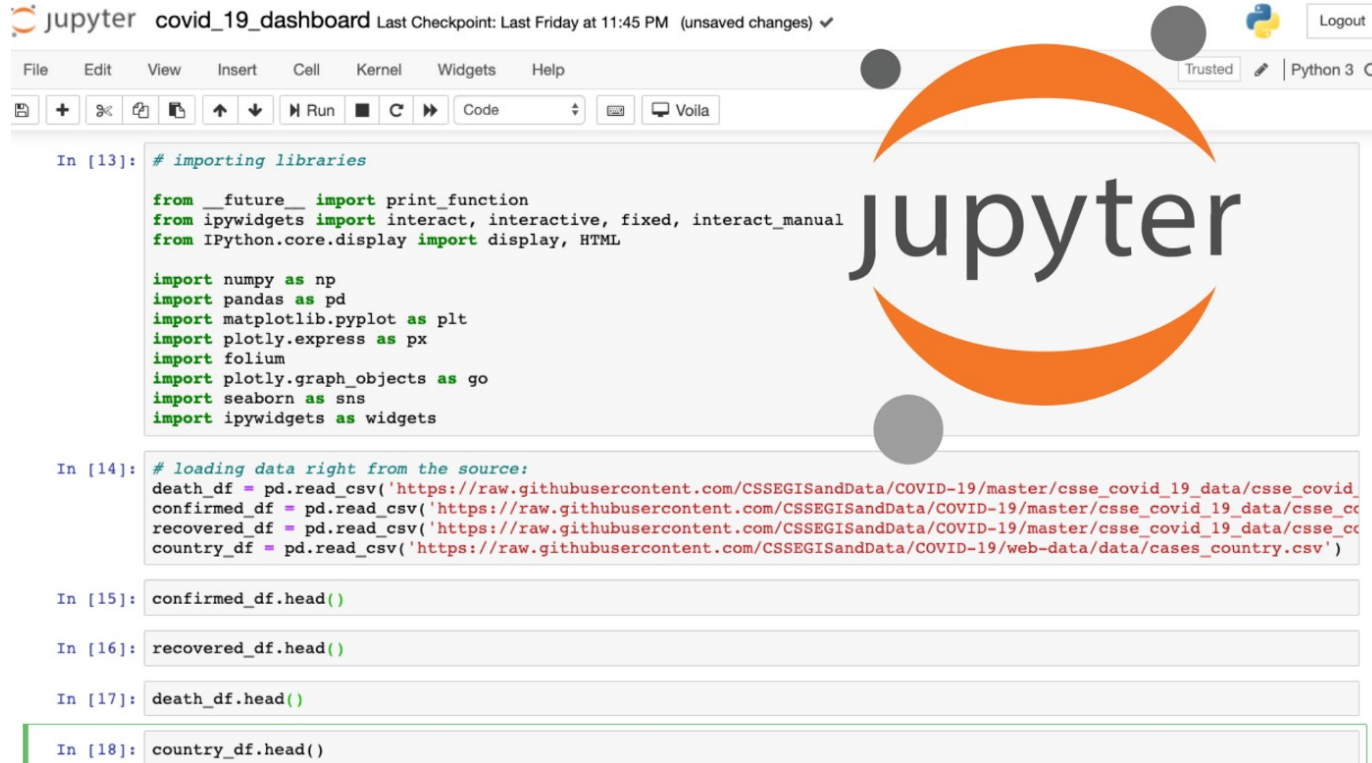
# R – Ross & Robert

- 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



# Jupyter Notebook



The image shows a Jupyter Notebook interface for a dashboard named 'covid\_19\_dashboard'. The top bar includes the Jupyter logo, the dashboard name, the last checkpoint time ('Last Friday at 11:45 PM'), and a status for 'unsaved changes'. A 'Logout' button is on the right. Below the top bar is a menu bar with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. A toolbar contains icons for file operations, running, and a 'Voilà' button. The notebook contains four code cells:

```
In [13]: # importing libraries

from __future__ import print_function
from ipywidgets import interact, interactive, fixed, interact_manual
from IPython.core.display import import display, HTML

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import folium
import plotly.graph_objects as go
import seaborn as sns
import ipywidgets as widgets

In [14]: # loading data right from the source:
death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')

In [15]: confirmed_df.head()

In [16]: recovered_df.head()

In [17]: death_df.head()

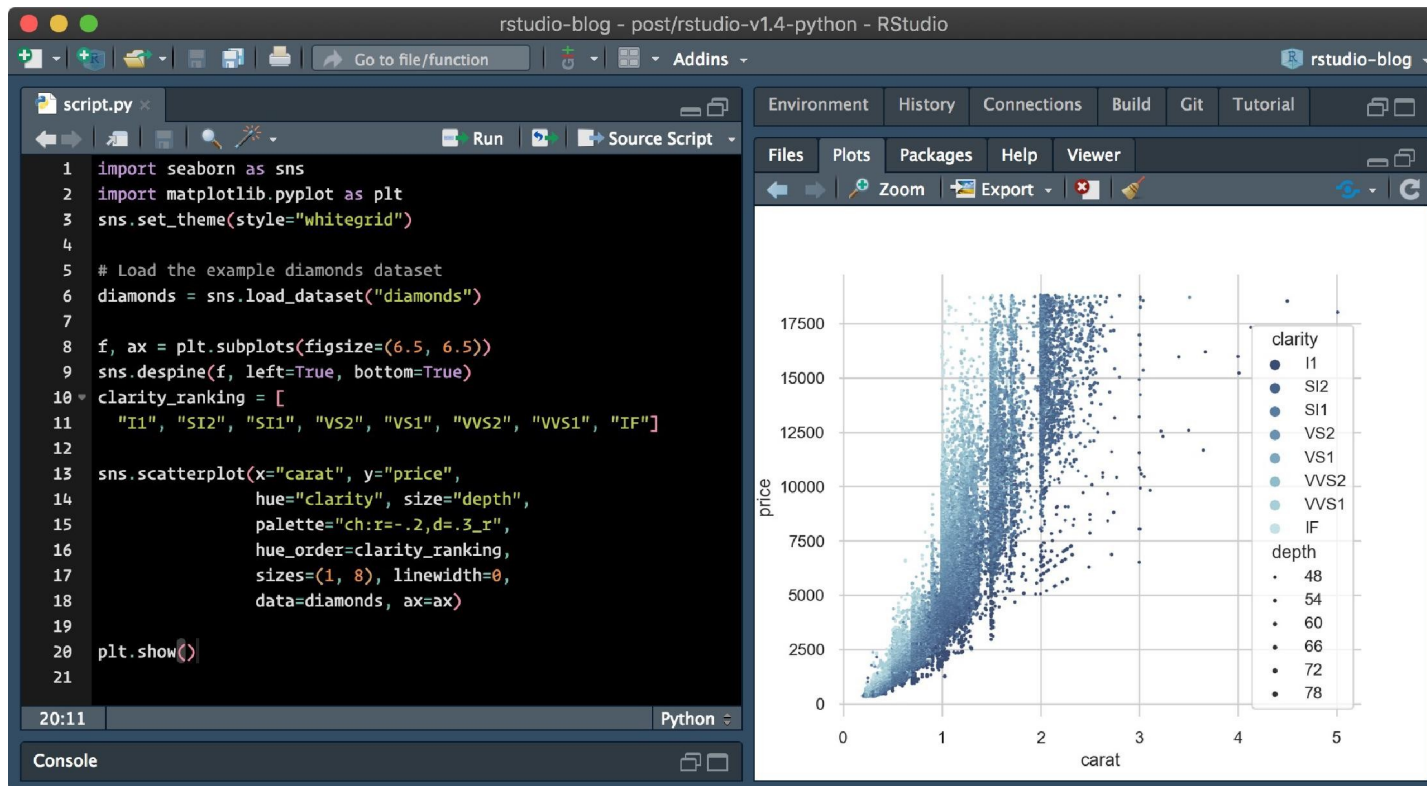
In [18]: country_df.head()
```



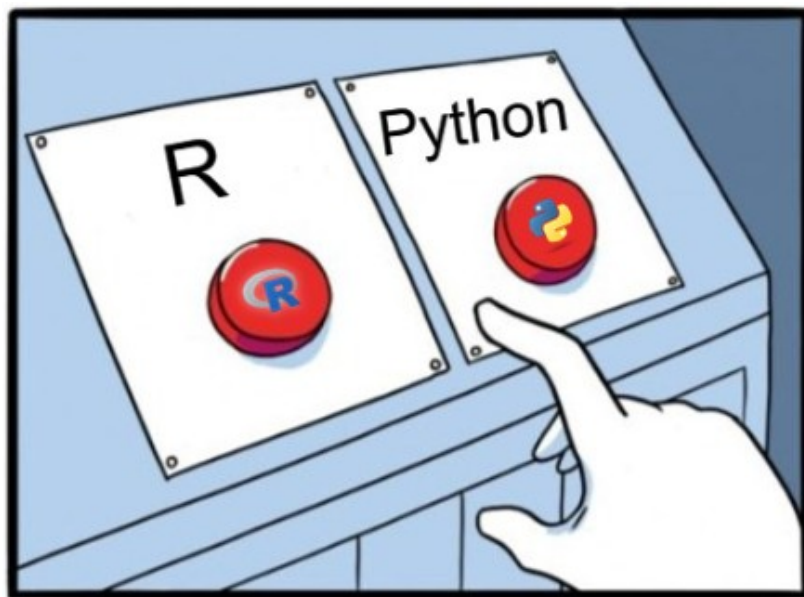
# Python

- 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.

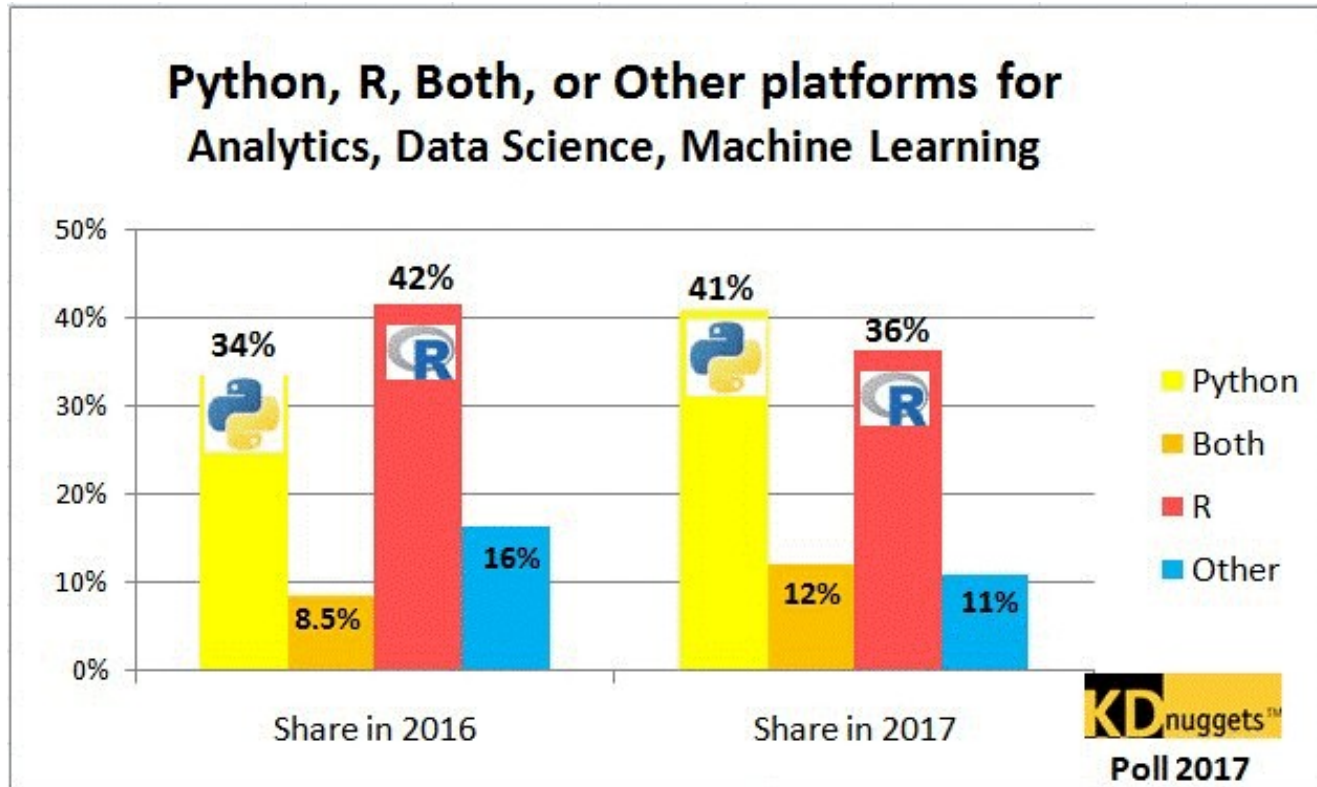
# Python e Código



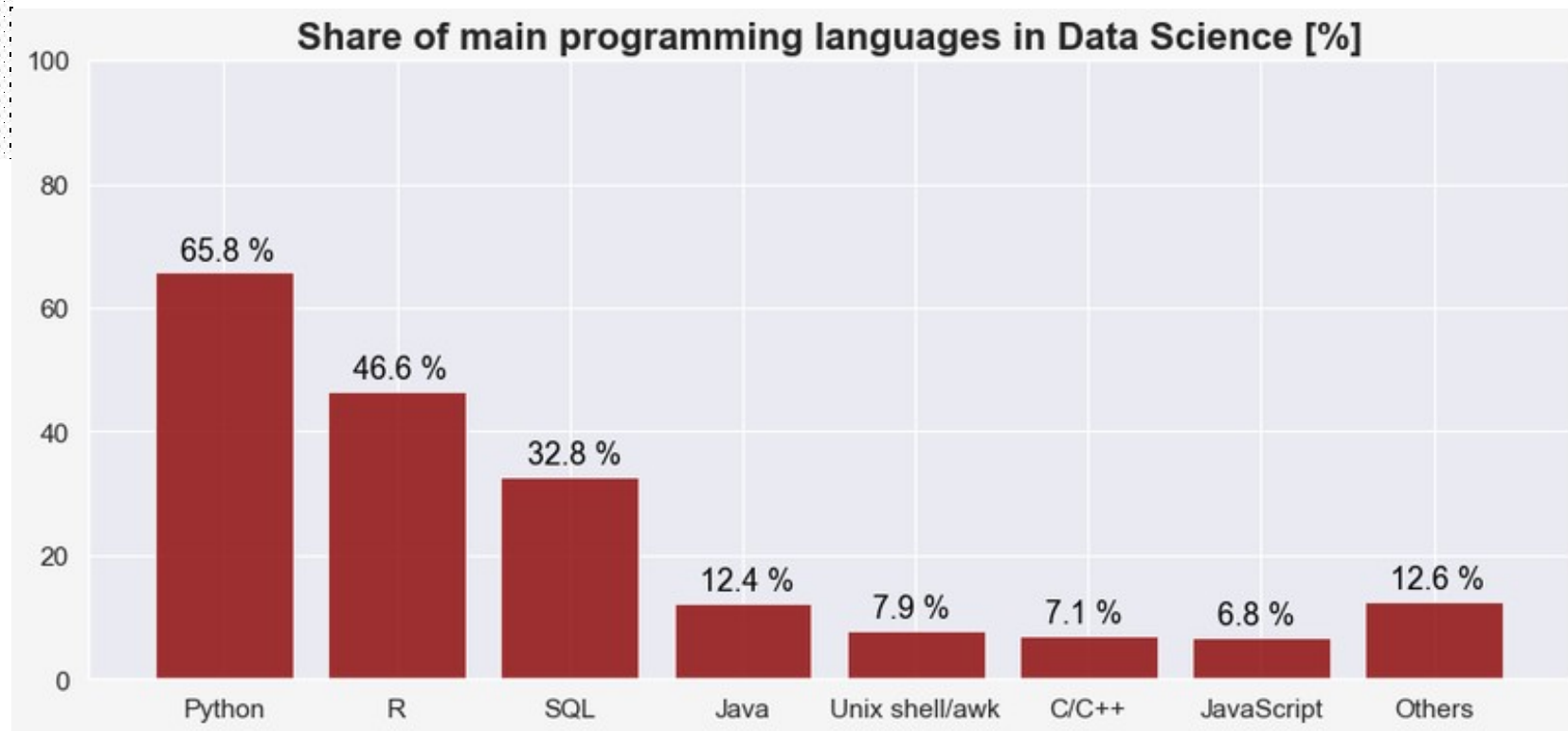
# R e Python em Ciência de Dados



# Cenário 2016/2017

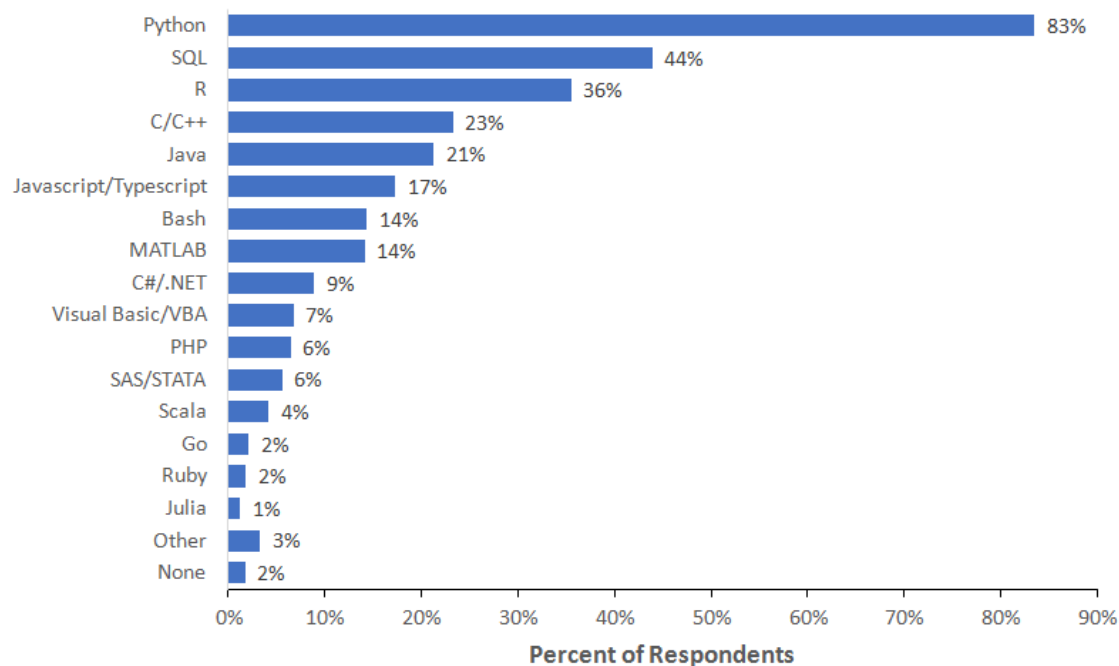


# Cenário 2019



# Reflexo no Kaggle

What programming language do you use on a regular basis?





# Então é Python?



Não.....talvez.



# Comparações gerais

Quesito		 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:

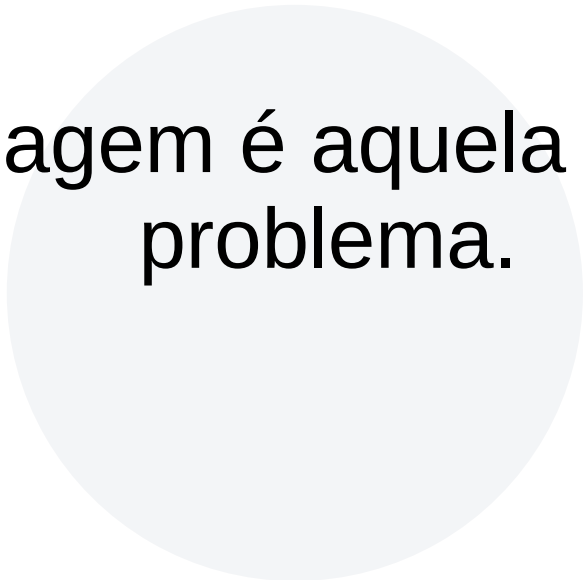


average Python enjoyer:

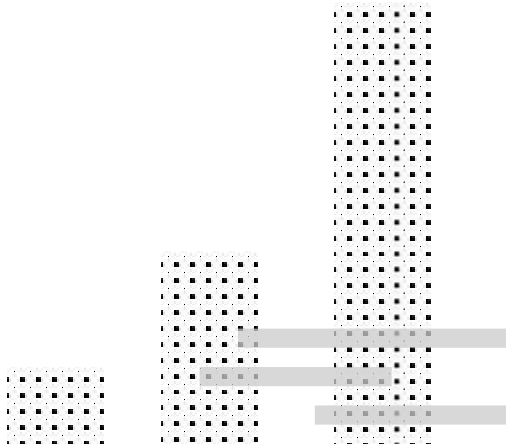




# Por que não ambos?




A melhor linguagem é aquela que resolve seu problema.





# Recomendações

- Google Colab;
  - PyCharm;
  - Spider.
- 
- 