

Aula 1: Ambientes de Programação

Prof. Mauricio Duarte

Linguagens...

Linguagens de programação mais utilizadas em Big Data (R e Python);

Coleta de dados, limpeza e integração.



Leituras recomendadas...

LINGUAGEM R – POR QUE É HORA DE APRENDER?

<http://datascienceacademy.com.br/blog/linguagem-r-por-que-e-hora-de-aprender/> (2018)

POR QUE CIENTISTAS DE DADOS ESCOLHEM PYTHON? (2019)

<http://www.cienciaedados.com/por-que-cientistas-de-dados-escolhem-python/>

R OU PYTHON PARA ANÁLISE DE DADOS?


<http://www.cienciaedados.com/r-ou-python-para-analise-de-dados/> (2019)

Gerenciador de aplicações

(<https://www.anaconda.com/distribution/>)

Anaconda Navigator

File Help

 ANACONDA NAVIGATOR

Home

Environments

Learning


Community

Documentation

Developer Blog

Twitter YouTube GitHub


Applications on Channels



JupyterLab
1.0.2

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.


Launch



Jupyter
Notebook
6.0.0

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

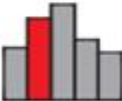
Launch



Spyder
3.3.6


Scientific PYTHON Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Launch




Glueviz
0.15.2

Multidimensional data visualization across files. Explore relationships within and among related datasets.



Orange 3
3.23.0

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

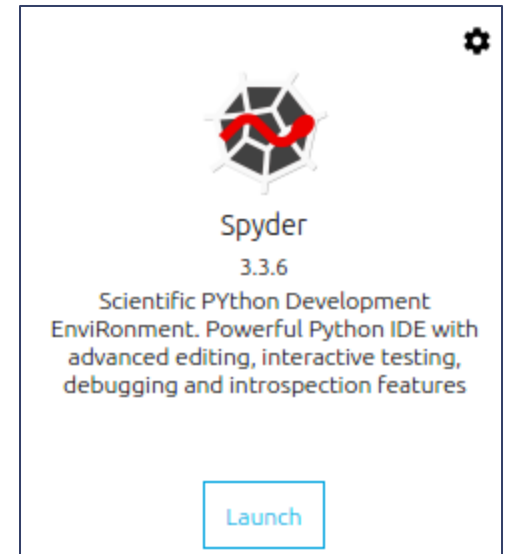


RStudio
1.1.456

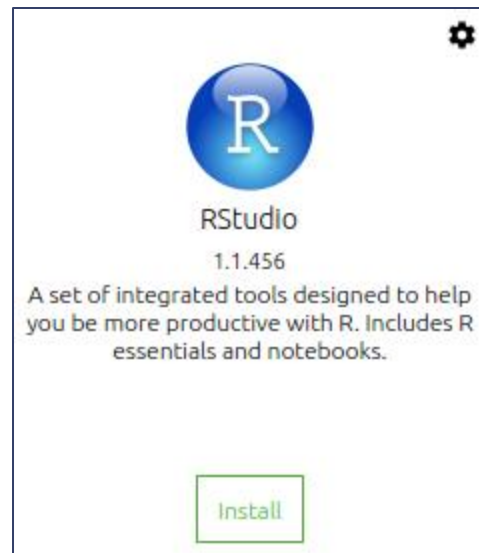
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Ambientes de programação

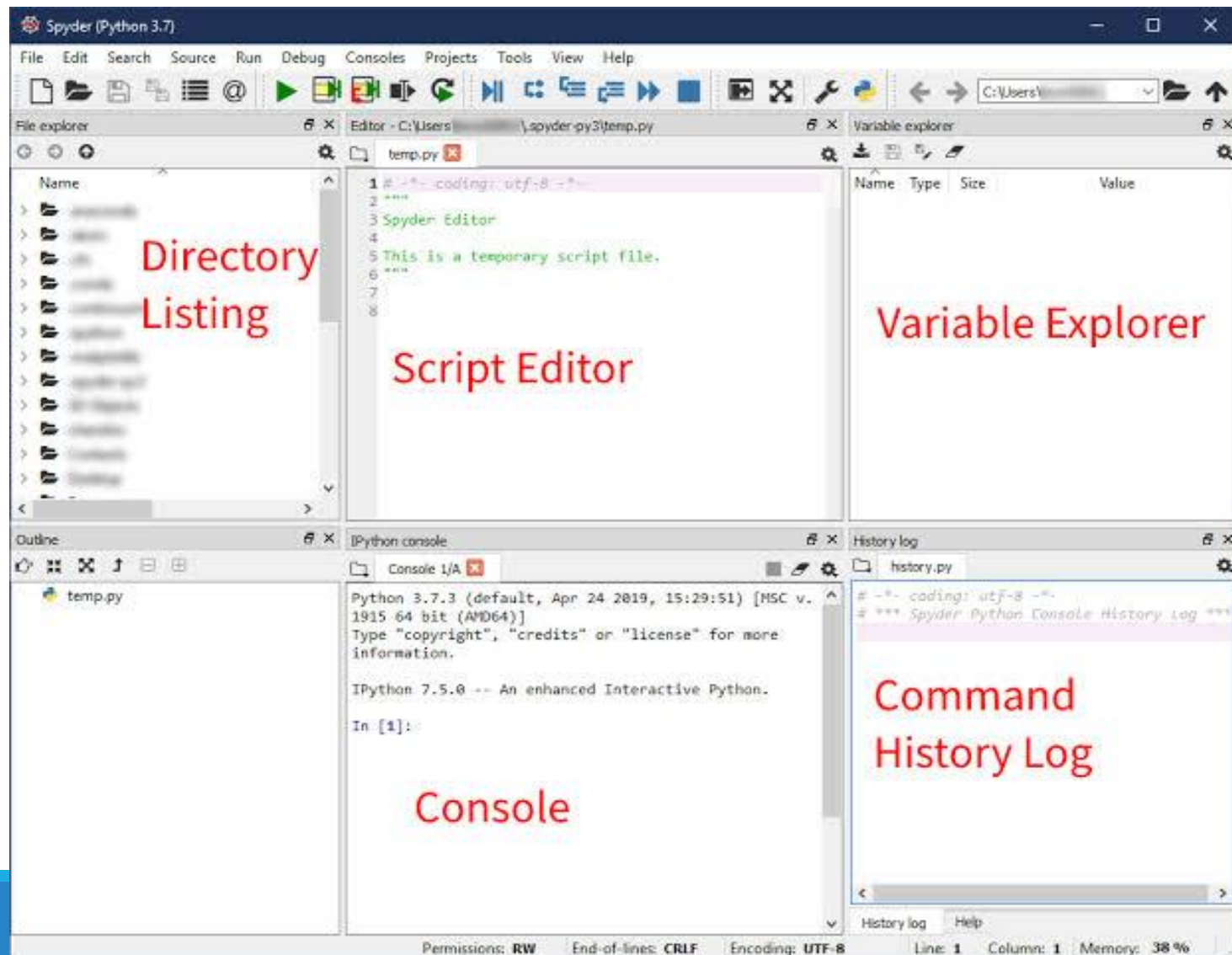
Anaconda - Spyder - Python



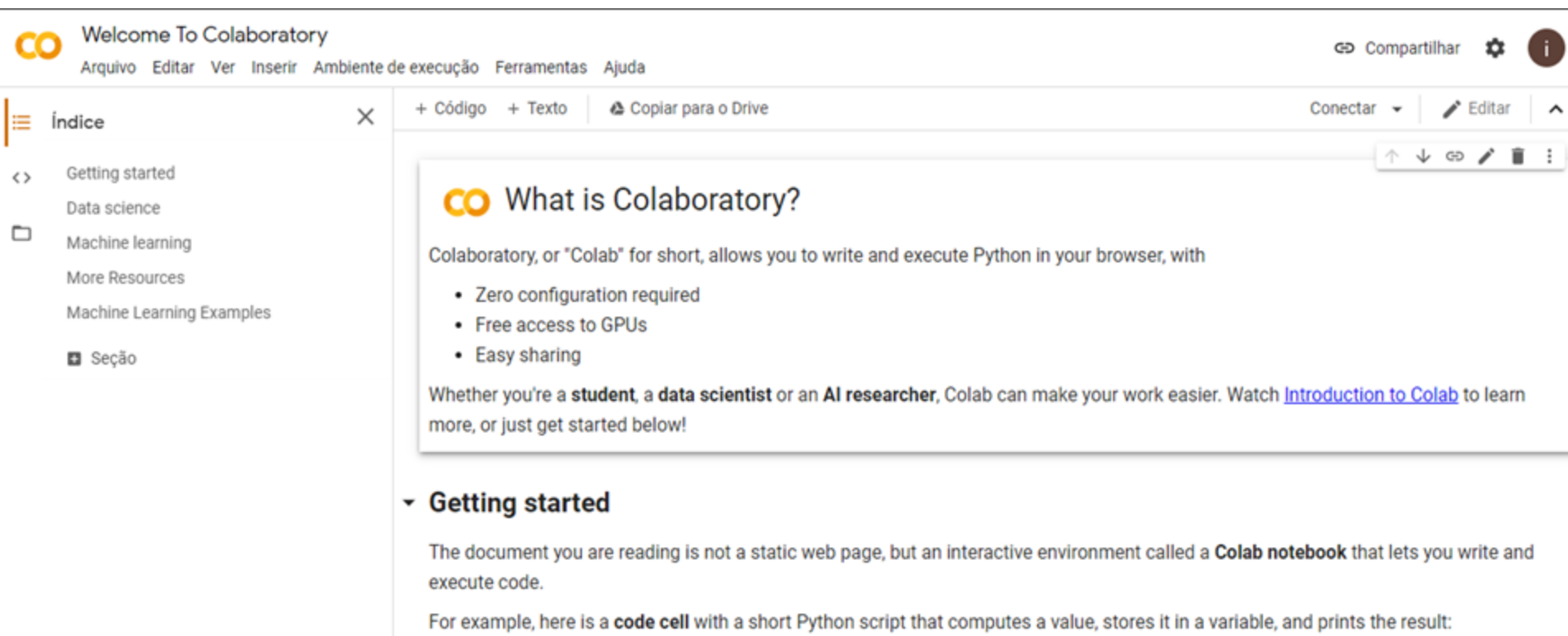
Anaconda - Rstudio - R



R Studio Online



Google Colab



The screenshot displays the Google Colaboratory web interface. At the top, the Google Colab logo is on the left, and the text 'Welcome To Colaboratory' is in the center. Below this, a navigation bar contains links: 'Arquivo', 'Editar', 'Ver', 'Inserir', 'Ambiente de execução', 'Ferramentas', and 'Ajuda'. On the right side of the top bar, there are icons for 'Compartilhar', settings, and a user profile. Below the navigation bar, a sidebar on the left shows a tree view under 'Índice' with items like 'Getting started', 'Data science', 'Machine learning', 'More Resources', and 'Machine Learning Examples'. The main content area is titled '+ Código + Texto' and 'Copiar para o Drive'. It features a section titled 'What is Colaboratory?' with a sub-header 'Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with'. This is followed by a bulleted list: 'Zero configuration required', 'Free access to GPUs', and 'Easy sharing'. Below the list, a paragraph states: 'Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch [Introduction to Colab](#) to learn more, or just get started below!'. The section is followed by a sub-header 'Getting started' and a paragraph: 'The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.' Below this, another paragraph says: 'For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:'.

Welcome To Colaboratory

Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda

Compartilhar

Índice

- Getting started
- Data science
- Machine learning
- More Resources
- Machine Learning Examples
- Seção

+ Código + Texto Copiar para o Drive

Conectar Editar

What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch [Introduction to Colab](#) to learn more, or just get started below!

Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

<https://www.alura.com.br/artigos/google-colab-o-que-e-e-como-usar>

Pandas

Pandas é uma biblioteca para manipulação e análise de dados, escrita em Python.

Essa é a biblioteca perfeita para iniciar suas análises exploratórias de dados.

Ela permite **ler**, **manipular**, **agregar** e **plotar** os dados em poucos passos.

<https://www.vooo.pro/insights/guia-de-acesso-rapido-ao-pandas/>

Exemplo Pandas no Google Colab

```
import pandas as pd

base_de_dados = pd.read_csv("https://raw.githubusercontent.com/alura-cursos/formacao-data-science/master/movies.csv")

print(base_de_dados)
```

```
movieId  ...  genres
0         1  ...  Adventure|Animation|Children|Comedy|Fantasy
1         2  ...           Adventure|Children|Fantasy
2         3  ...           Comedy|Romance
3         4  ...       Comedy|Drama|Romance
4         5  ...           Comedy
...      ...  ...
9737    193581  ...  Action|Animation|Comedy|Fantasy
9738    193583  ...       Animation|Comedy|Fantasy
9739    193585  ...           Drama
9740    193587  ...       Action|Animation
9741    193609  ...           Comedy
```

[9742 rows x 3 columns]

SciKit-sklearn

Scikit-sklearn é uma biblioteca Python amplamente usada para projetos que envolvem aprendizado de máquina.

Bases de Dados em Agricultura

- **Genbank** (<https://www.ncbi.nlm.nih.gov/genbank/>), o banco de dados de sequências genéticas, uma coleção anotada de todas as sequências de DNA disponíveis ao público;
- **Base de Dados de Pesquisa Agropecuária** (<https://www.bdpa.cnptia.embrapa.br/consulta/busca>);

GenBank....

National Center for Biotechnology Information



Some NLM-NCBI services and products are experiencing heavy traffic, which may affect performance and availability. We apologize for the inconvenience and appreciate your patience. For assistance, please contact our Help Desk at info@ncbi.nlm.nih.gov.

Nucleotide

Nucleotide ▾

Citrus x limon plastid |



Search

[Create alert](#) [Advanced](#)

Species

[clear](#)

Summary ▾ 20 per page ▾ Sort by Default order ▾

Send to: ▾

Filters: [Manage Filters](#)

Plants (8)

[Customize ...](#)

Molecule types

genomic DNA/RNA (7)

Analyze these sequences

[Run BLAST](#)

Citrus x limon plastid

Organismo Modelo



Na pesquisa.... 1º. Link...

 Filters activated: Plants. [Clear all](#)

☐ [Citrus x limon plastid, complete genome](#)

1. 160,101 bp circular DNA

Accession: KY085897.1 GI: 1184801769

[BioProject](#) [Protein](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)



Baixando a base de dados FASTA...

FASTA ▾

Citrus x limon plastid, complete genome

GenBank: KY085897.1

[GenBank](#) [Graphics](#)

>KY085897.1 Citrus x limon plastid, complete genome

```
TGGGCGAACGACGGGAATTGAACCCGCGCATGGTGGATTACAATCCACTGCCTTGATCCACTTGGCTAC
ATCCGCCCCCTCCGCTATTTACACAATTTTGAATACAAAGATCTAAAATCAACCATTGATTATTTTTGT
TTATCTTATCTTACTTATGAAGAGCCAAATGAAGATCGAAGAGCAGAAAACATAACCTTTCTATTGTCT
TTTTTCTTTGCTATGAAATTAAGTGTAAATAGAACTAATTTCTAATTAATAATCTAATAATAAAATT
AGAAATTTAGTAATTTATTAGTAGTATTAGCGCATACCAACAATATCATACTAAATCAAAGAAAAGAAAAG
CATAAAATACTTAACAAAAAATGAAGTAAAACTAATAAAGAACCCCGATAAGAAACCC
GACTAAATAACGGATCAATACTGACGGGTCAGTATTGATCCGTTATTTTCAAAAACCCGCTACACAAAG
ACCAAAATCTTACCCATTTGTAGATGGGGCTTCAATAGCAGCTAGGTCTAGAGGGGAAGTTATGAGCATT
CGTTCATGCATAACTTCCATACCAAGGTTAGCACGATTAATAATATCAGCCAGGTATTAATTACACGAC
CTTGACTATCAACTACAGATTGGTTGAAATTGAAACCATTTAAGTTGAAAGCCATAGTGCTAATACCTAA
AGCAGTGAACAGATACCTACTACAGGCCAAGCAGCCAGGAAGAAATGTAAGAACGAGAATTGTTGAAA
CTAGCATATTGGAAGATCAATCGGCCAAAATAACCGTGAGCAGCTACGATATTATAAGTTTCTTCCTCTT
GACCGAATCTGTAACCTGCATTAGCAGATTCTTTTCTGTGGTTTCCCTGATCAAACTAGAGGTTACCAA
GGAACCATGCATAGCACTGAATAGGGAGCCGCCGAATACACCAGCTACGCCTAACATGTGGAATGGGTGC
ATAAGGATGTTGTGCTCAGCCTGGAATACAATCATGAAATTGAAAGTACCAGAGATTCTAGAGGCATAC
CATCAGAAAACTTCTTGACCGATTGGGTAGATCAAGAAAACAGCAGTCGCTGCTGCAACAGGAGCTGA
ATATGCAACAGCAATCCAAGGACGCATACCCAGACGGAACTAAGTTCCTCACTCACGCCCATGTAACAA
GCTACACCAAGTAAGAAGGTAGAACAAATTAGCTCATAAGGACCGCCATTGTATAACCATTCATCAACGG
ATGCCGCTTCCCATATCGGGTAAAAATGCAAACTATAGCTGCAGAAAGTAGGAATAATCGCACCAGAAAT
```

Send to: ▾

☒ Complete Record

☐ Coding Sequences

☐ Gene Features

Choose Destination

☒ File

☐ Clipboard

☐ Collections

☐ Analysis Tool

Download 1 item.

Format

FASTA ▾

Show GI ☐

Create File

Gene

PubMed (Weighted)

LinkOut to external resources

Dravid Digital Repository

Atividade prática

Objetivos:

- 1.) Contar a quantidade de “A” (Adenina) ; “C” (Citosina); “T” (Timina) e “G” (Guanina)
- 2.) Emitir o percentual médio de cada uma delas.

Leituras Complementares...

Python - código para ler um arquivo no formato FASTA e transformá-lo em lista.

Disponível em:

<https://gist.github.com/marcoscastro/89e8c66703d5067b9b3c>

Trabalhando com Arquivos. Disponível em:

<https://panda.ime.usp.br/pensepy/static/pensepy/10-Arquivos/files.html>

Ler arquivos fasta no python e ignorar a primeira linha

Disponível em:

<https://pt.stackoverflow.com/questions/236391/ler-arquivos-fasta-no-python-e-ignorar-a-primeira-linha/>

Atividade Avaliativa – Aula 1

- Pesquisar como trabalhar com as bibliotecas pandas no uso de funções matemáticas básicas (média, mediana, moda e desvio padrão). Crie um pequeno guia de usuário. (Isso não é para entregar, ok).
- Faça um programa no Colab, que crie/gere uma lista capaz de armazenar 1000 idades de pessoas (valores entre 0 a 100). Mostre a idade media, a moda e a mediana. Considerando agora, que precisaremos ter apenas idades de adultos (≥ 21), faça com que todas as idades menores que 21 sejam substituídas pela media das idades ≥ 21 anos.