

Aula 2: 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.

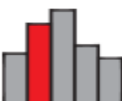
Launch



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.

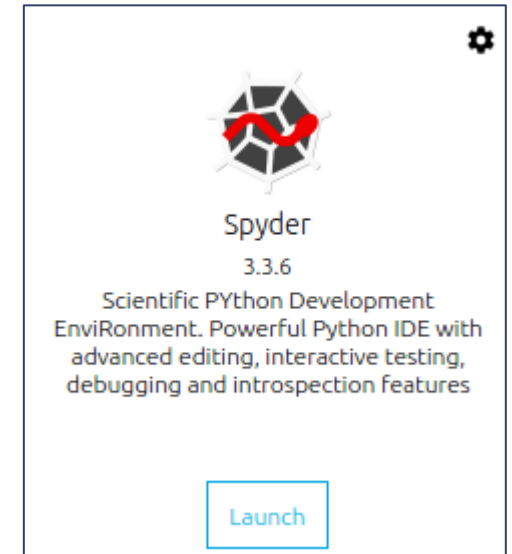


RStudio
1.1.456

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

Ambientes de programação

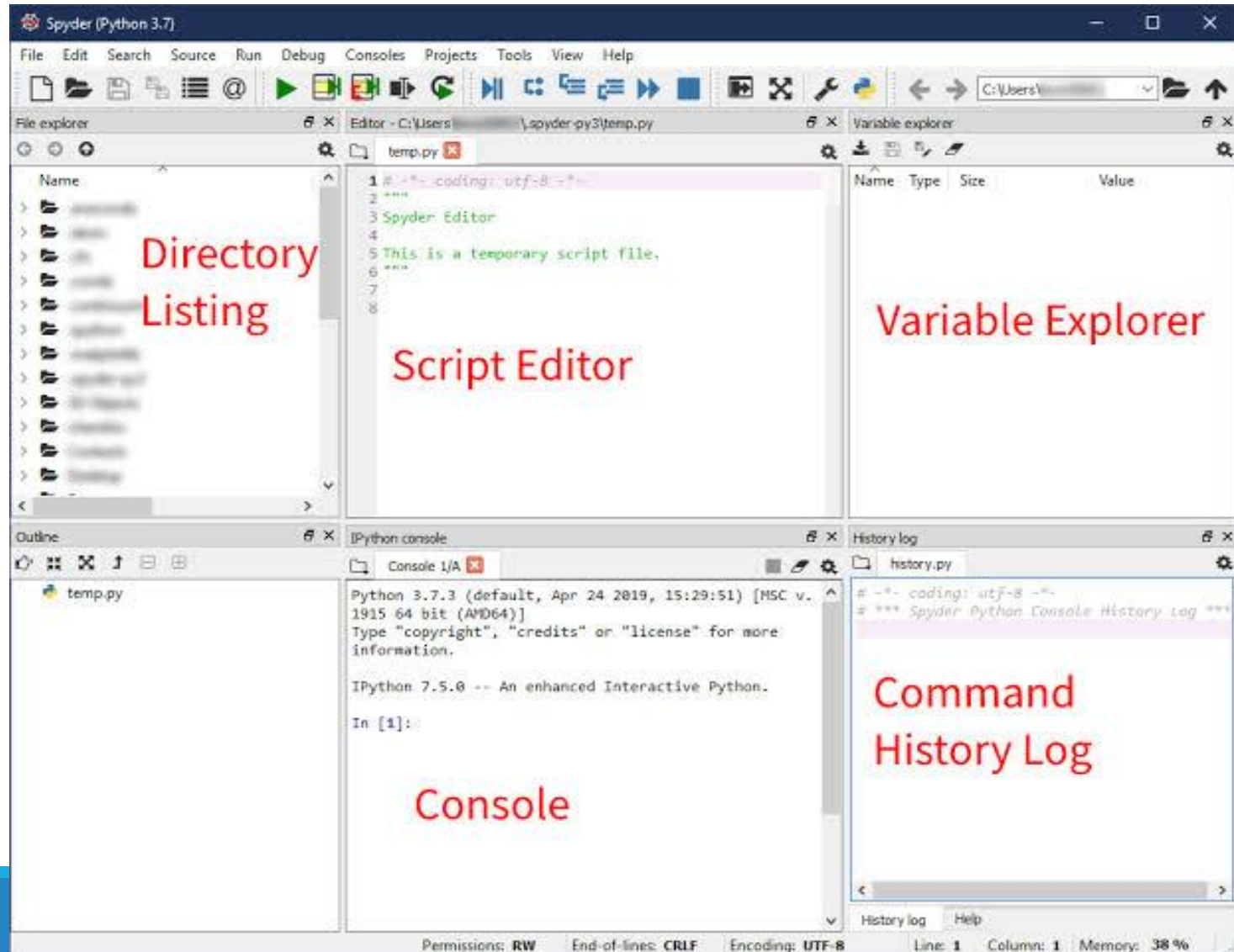
Anaconda - Spyder - Python



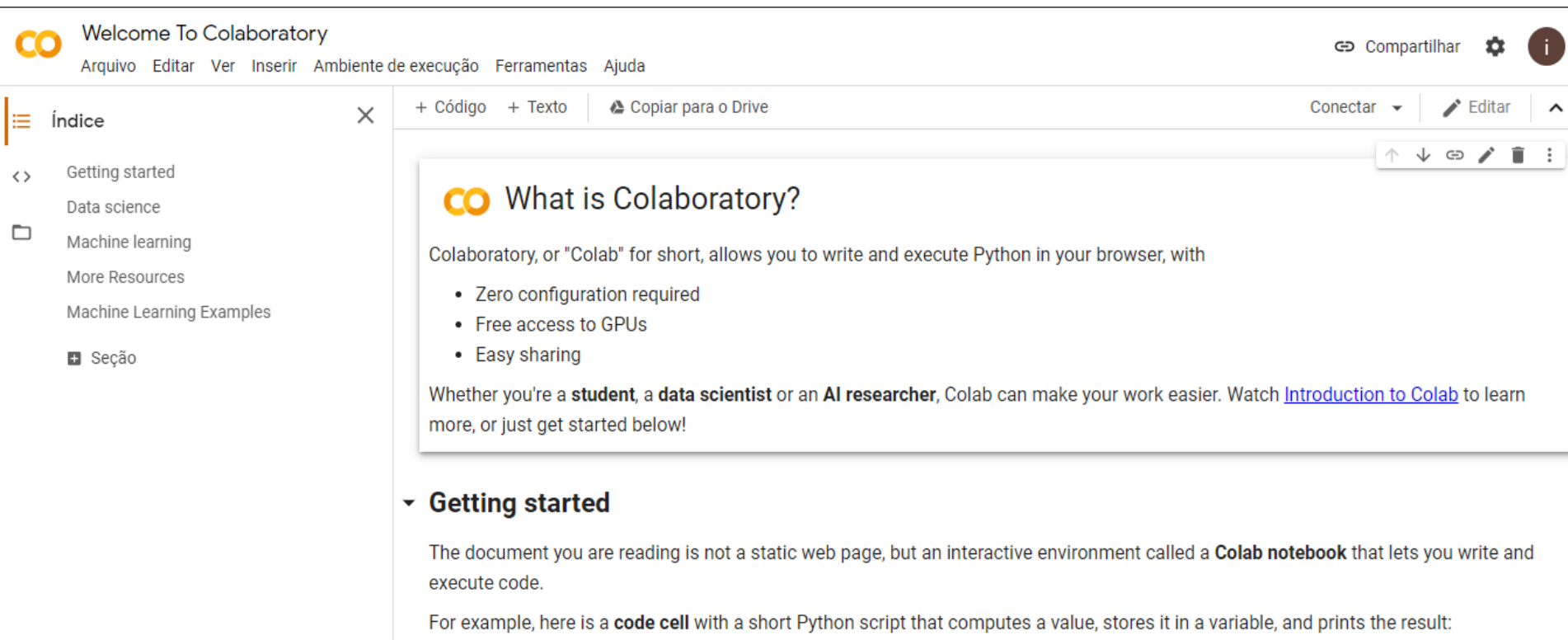
Anaconda - Rstudio - R



Ambiente Spyder



Google Colab



The screenshot displays the Google Colaboratory web interface. At the top, there's a header with the Colab logo and the text 'Welcome To Colaboratory'. Below this, a navigation bar includes links for 'Arquivo', 'Editar', 'Ver', 'Inserir', 'Ambiente de execução', 'Ferramentas', and 'Ajuda'. On the right side of the header, there are icons for 'Compartilhar', settings, and a user profile. A sidebar on the left contains an 'Índice' (Index) section with links to 'Getting started', 'Data science', 'Machine learning', 'More Resources', and 'Machine Learning Examples'. The main content area is titled '+ Código' and '+ Texto', with a 'Copiar para o Drive' button. It features a section titled 'What is Colaboratory?' which explains that Colab allows writing and executing Python in a browser. This section lists three key features: 'Zero configuration required', 'Free access to GPUs', and 'Easy sharing'. It also mentions that Colab is suitable for students, data scientists, and AI researchers, and provides a link to 'Introduction to Colab'. Below this, a 'Getting started' section begins by stating that the document is an interactive 'Colab notebook' and provides an example of a 'code cell' containing a short Python script.

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

+ Código + Texto Copiar para o Drive

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....

→ ↻ ncbi.nlm.nih.gov/genbank/

NCBI Resources ▾ How To ▾

GenBank

GenBank ▾ Submit ▾ Genomes ▾ WGS ▾ Metagenomes ▾ TPA ▾ TSA ▾ INSDC ▾ Other ▾

Data regarding the SARS-CoV-2 (2019-nCoV, Wuhan coronavirus) outbreak sequences can be found in [GenBank/SRA](#), the [NCBI Virus](#) resource, and a specialized [BLAST](#) page that searches Betacoronavirus sequences.

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan;41\(D1\):D36-42](#)). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using [NCBI e-utils](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)

Arabidopsis thaliana

Genoma

Organismo Modelo



Na pesquisa.... 1º. Link...

☒ [Arabidopsis thaliana chromosome 1 sequence](#)

1. 30,427,671 bp linear DNA

Accession: CP002684.1 GI: 332189094

[Assembly](#) [BioProject](#) [BioSample](#) [Protein](#) [PubMed](#) [Taxonomy](#)

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



Nucleotide

Nucleotide ▼

Advanced

Search

Fasta

FASTA ▼

Send to: ▼

Arabidopsis thaliana chromosome 1 sequence

GenBank: CP002684.1

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

>CP002684.1 Arabidopsis thaliana chromosome 1 sequence

```
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAATCTTTAAATCC
TACATCCATGAATCCCTAAATACCTAATTCCTAAACCCGAAACCGGTTTCTCTGGTTGAAATCATTGT
GTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTATTGTTGTGTGTAGATTTTTTAAAAATATCA
TTTGAGGTCAATACAAATCCTATTTCTTGTTGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTC
ATTTGTTATATTGGATACAAGCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTGTAACCTTA
GGGTTGGTTTATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTTATTGTCATTCTT
ACTCCTTTGTGGAATGTTTGTCTATCAATTTATCTTTTGTGGGAAAATTATTTAGTTGTAGGGATGAA
GTCTTTCTTCTGTTGTTGTTACGCTTGTCTCATCTCTCAATGATATGGGATGGTCCTTTAGCATTAT
TCTGAAGTTCTTCTGCTTGATGATTTTATCCTTAGCCAAAAGGATTGGTGGTTTGAAGACACATCATATC
AAAAAAGCTATCGCCTCGACGATGCTCTATTTCTATCCTTGTAGCACACATTTTGGCACTCAAAAAAGTA
TTTTTAGATGTTTGTGTTTCTTCTTGAAGTAGTTTCTCTTGTGAAAATTCCTCTTTTTTTAGAGTGATT
TGGATGATTCAAGACTTCTCGGTACTGCAAAGTTCTTCCGCCTGATTAATTATCCATTTTACCTTTGTCTG
TAGATATTAGGTAATCTGTAAGTCAACTCATATACAACCTATAATTTAAAATAAAATTATGATCGACACA
CGTTTACACATAAAATCTGTAATCAACTCATATACCCGTTATTTCCACAATCATATGCTTTCTAAAAGC
AAAAGTATATGTCAACAATTGGTTATAAATTATTAGAAGTTTTCCACTTATGACTTAAGAACTTGGAAG
CAGAAAGTGGCAACACCCCCCACCTCCCCCCCCCCCCCCCCCAAAATTGAGAAGTCAATTTTATAT
AATTTAATCAAATAAATAAGTTTATGGTTAAGAGTTTTTACTCTCTTATTTTTCTTTTTCTTTTGTGAG
ACATACTGAAAAAAGTTGTAATTATTAATGATAGTTCTGTGATTCCTCCATGAATCACATCTGCTTGATT
TTTCTTTCATAAATTTATAAGTAATACATTCTTATAAAATGGTCAGAGAAACACCAAAGATCCCAGATT
TCTTCTCACTTACTTTTTTCTATCTATCTAGATTATATAAATGAGATGTTGAATTAGAGGAACCTTTGA
TTCAATGATCATAGAAAAATTAGGTAAGAGTCAAGTGTGTTATGTTATGGAAGATGTGAATGAAGTTTG
ACTTCTCATTGTATATGAGTAAAATCTTTTCTTACAAGGGAAGTCCCAATTGGTCAACATGTGAAAGCA
CGTGTCTATGTTCTTACTTTTGTGTTGGGAATCTTCTAATTACTGTATATGGAAGATGTGAATGAAGTTT
GGTCTGAATGTGGCCAAGGTTCCGTCATTTGGAGATACGAAATCAAATCTCCTTTAAGATTTTGTGTTT
ATAATGTGTTCTTCCATCCACATCTATCTCCATATGATATGGACCATATCATACATCATCATTGTGCA
```

Downloading Large Sequence: 20.63MB

- ☒ Complete Record
☐ Coding Sequences
☐ Gene Features

Choose Destination

- ☒ File ☐ Clipboard
☐ Collections ☐ Analysis Tool

Download 1 item.

Format

FASTA ▼

Show GI ☐

Create File

[BioProject](#)[BioSample](#)[Protein](#)[PubMed](#)[Taxonomy](#)[Component Of](#)[Full text in PMC](#)[Gene](#)

Atividade prática

Objetivos:

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

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

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

Leituras Complementares...

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 de pesquisa

- 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.
- Faça testes no Colab para entender o funcionamento de tais as funções.