



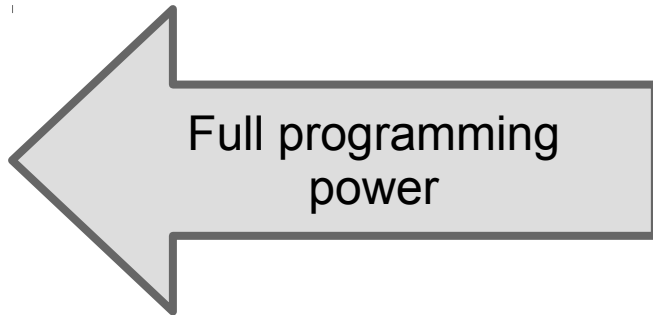
Ciência de Dados Python - Ferramentas

Luiz Celso Gomes-Jr

Ferramentas

- Linguagens (Python, R, MATLAB...)
- Ambientes/Distribuições (Anaconda)
- Pacotes (pandas, scikit-learn, SciPy...)
- IDEs (spyder, R-Studio)

Linguagens/Ambientes



Business Insider

What it's like to have the best job in America right now



Dominic Umbro

🕒 1h 🔥 35,221



Employees say being a data scientist is the best job in the US. [Glassdoor](#)

[http://www.businessinsider.com/data-scientist-best-job-in-us-right-now-2018-](http://www.businessinsider.com/data-scientist-best-job-in-us-right-now-2018-2)

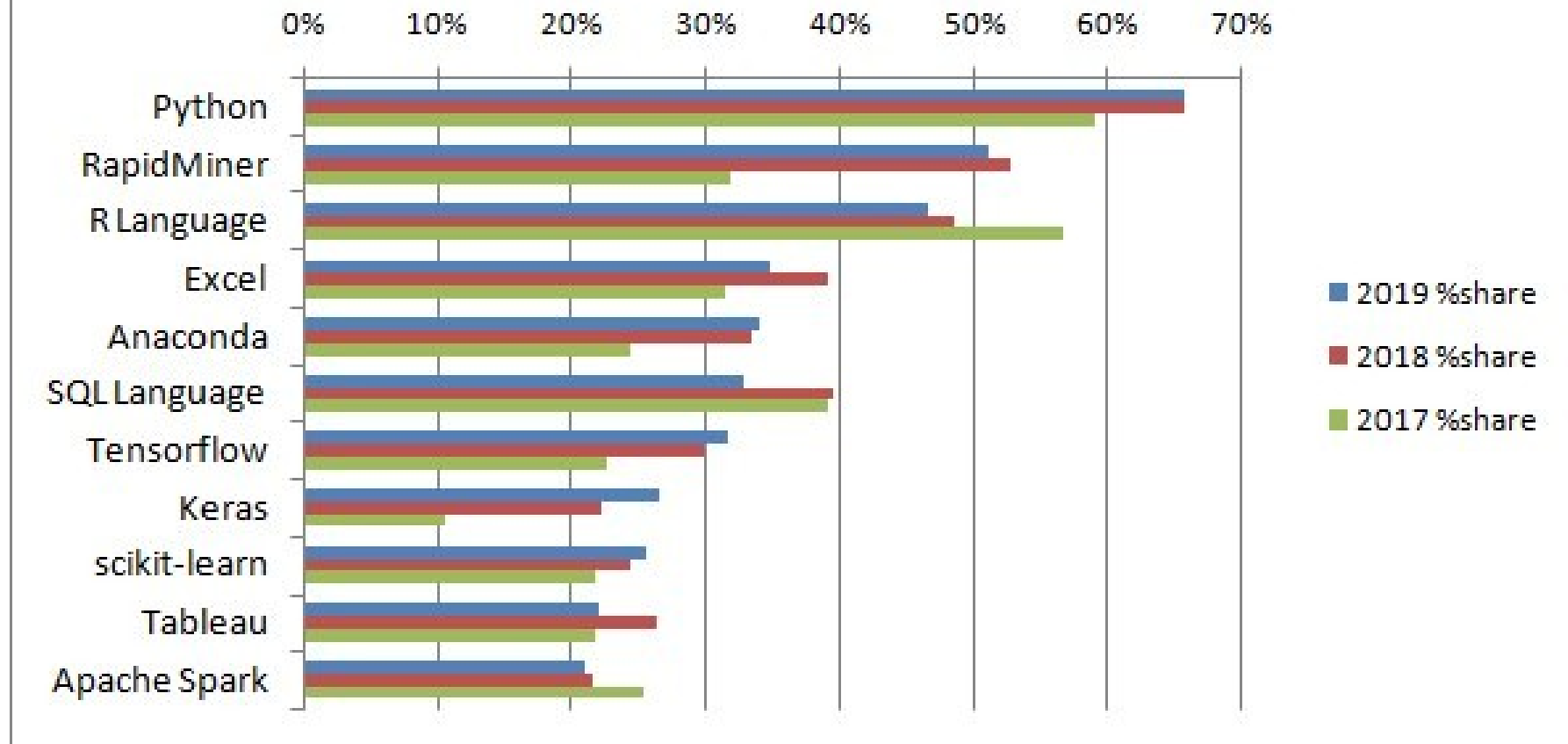
Business Insider - skills

- SQL
- Python/R (maybe Tableau)
- Modeling
- Dashboarding (visualization)
- More recent big data tools like Hadoop

"Just as important is having product and business sense," Cheng said. "Having your own **intuition and understanding** of your **subject area** so that you can come up with **good questions** and build **models that actually make sense** to the area that you're building."

Market Share




















Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll





Programming language popularity: Python tightens its grip at the top

Python retains its top spot as the most popular language for electrical and electronics engineers.

Rank	Language	Type	Score
1	Python	  	100.0
2	Java	  	96.3
3	C	  	94.4
4	C++	  	87.5
5	R		81.5
6	JavaScript		79.4
7	C#	   	74.5
8	Matlab		70.6

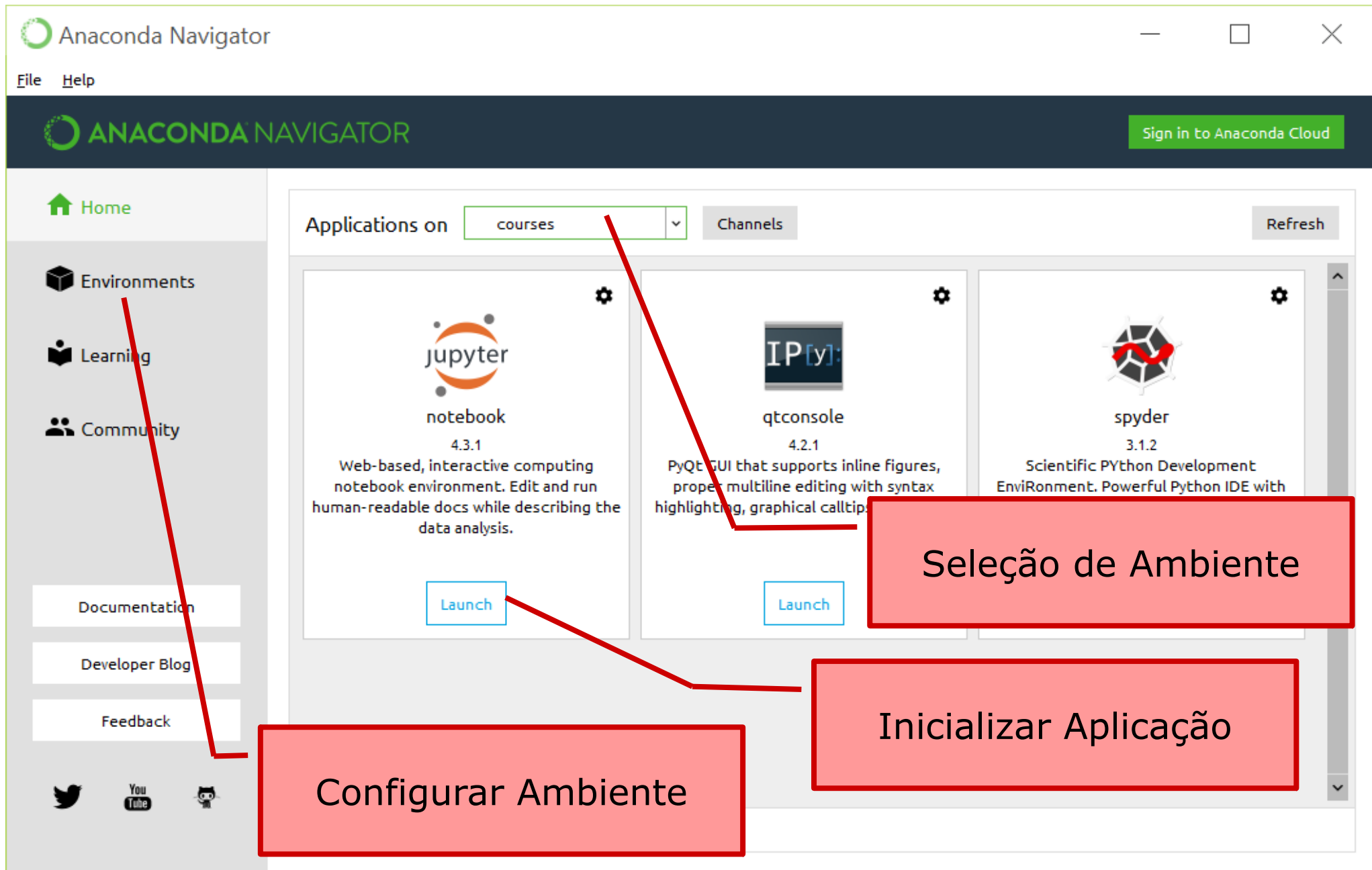




Anaconda

- Anaconda é uma distribuição **open source** para as linguagens **Python** e **R**. Oferece opções para processamento em larga escala, análise preditiva, computação científica, etc. com o objetivo de simplificar o **gerenciamento de pacotes**.
- Usa **conda** para gerenciamento de pacotes

Anaconda Navigator



Anaconda Navigator

The screenshot shows the Anaconda Navigator application window. The left sidebar contains navigation links: Home, Environments, Learning, and Community. The main area displays a list of environments, with 'courses' selected. A table of installed packages is shown for the 'courses' environment. Two red boxes with arrows point to the 'Create' button and the package list. A third red box points to the 'courses' environment in the list.

Anaconda Navigator

File Help

ANACONDA NAVIGATOR

Sign in to Anaconda Cloud

Home

Environments

Learning

Community

Documentation

Developer Blog

Feedback

Twitter YouTube GitHub

Create Clone Import Remove

Search Environments

root

courses

Installed Channels Update index... pan

Name	T	Description	Version
✓ pandas		Powerful data structures and data analysis tools	0.19.2
✓ pandas-datareader		Data readers extracted from the pandas codebase	0.2.1
✓ pandasql		SqlDf for pandas	0.7.3
✓ pandocfilters			
✓ qgrid			

5 packages available matching "pan" (C:\Users\Luiz\Anaconda2\envs\courses)

Instalar pacotes

Novos ambientes



Jupyter Notebook

- É uma ferramenta que unifica a escrita de código, sua documentação e a visualização dos resultados.
- Usado para limpeza e transformação de dados Data, simulação numérica, modelagem estatística, aprendizado de máquina, etc.
- Inicialmente focado em Python, mas já suporta mais de 40 linguagens como **R**, Julia e Scala.

Jupyter Notebook

The image shows a web browser window displaying the Jupyter Notebook interface. The browser's address bar shows the URL `localhost:8889/tree/Dropbox/utfpr/teaching/DataScience/`. The Jupyter logo is visible in the top left, and a 'Logout' button is in the top right. Below the logo, there are tabs for 'Files', 'Running', and 'Clusters'. The 'Files' tab is active, showing a file tree. A red box labeled 'Notebooks' points to the 'New' button in the top right of the file tree. Another red box labeled 'Novo notebook' points to the 'Explore156.ipynb' file in the list. The file list shows two files: 'Explore156.ipynb' and 'Read156.ipynb', both with a green 'Running' status indicator.

Luiz Celso

localhost:8889/tree/Dropbox/utfpr/teaching/DataScience/

Apps Inbox Keep Temp Personal PhD UTFPR WhatsApp Other bookmarks

jupyter Logout

Files Running Clusters

Select items to perform actions on them.

Upload New

Dropbox / utfpr / teaching / DataScience / 201701 / lessons / 03 - Exploratory Analysis

..

Explore156.ipynb Running

Read156.ipynb Running

Notebooks

Novo notebook

Jupyter Notebook

The screenshot displays a Jupyter Notebook interface in a web browser. The browser's address bar shows the URL: `localhost:8888/notebooks/Dropbox/utfpr/teaching/DataScience/201701/lessons/03%20-%2`. The notebook's title bar indicates the file name is `aluguel` and it is autosaved. The interface includes a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. Below the menu bar is a toolbar with icons for saving, creating new cells, deleting cells, moving cells, and running code. The main content area shows a code cell with the following Python code:

```
In [64]: fig = plt.figure()
ax = fig.add_subplot(111)
df.plot.scatter(x='area', y='aluguel', ax=ax)
plt.show()
```

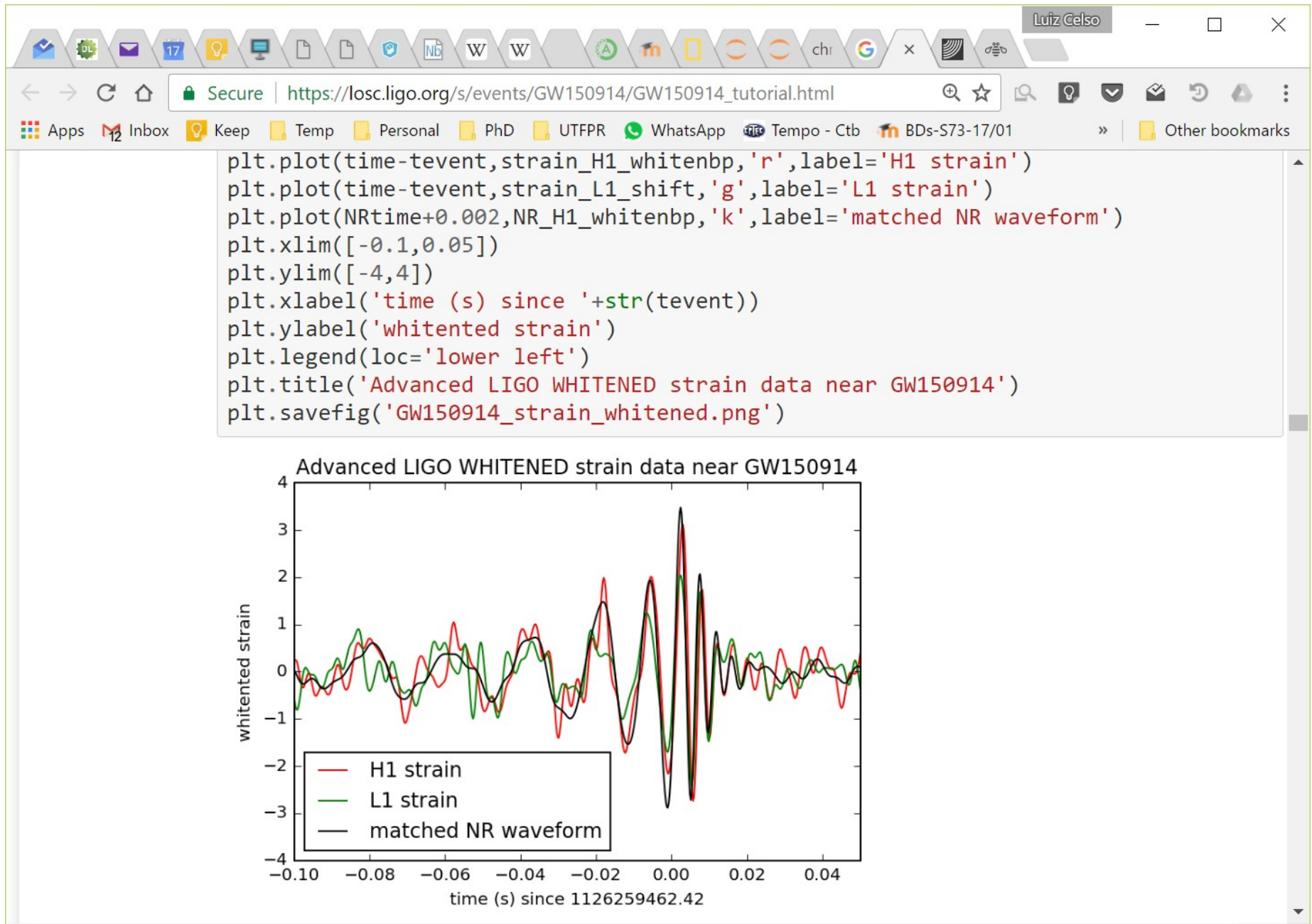
Below the code cell is a scatter plot titled "Visualizando a correlação entre as v". The y-axis is labeled "aluguel" and ranges from 600 to 950. The x-axis is labeled "area". The plot shows a positive correlation between the area and the rental price. Two red boxes with arrows point to specific features: one points to the "Cell" menu and the "Run" button in the toolbar, labeled "Gerenciamento de células"; the other points to the "Run" button, labeled "Executar código".

Visualizando a correlação entre as v

Gerenciamento de células

Executar código

LIGO + Jupyter



Documentação/Código

Formatação

É possível formatar a célula de documentação de diversas formas. Para ver como esta célula que você está lendo foi formatada, clique duas vezes para iniciar a edição. A linguagem usada para fazer a formatação é chamada *Markdown*. Ela permite diversas coisas como:

- Listas
- *Itálico*
- **Negrito**
- Fórmulas matemáticas: $f(x) = x^2 + 2x + \log_2 x$
- Links: [UTFPR](#)

Para conhecer outras possibilidades de uso de *Markdown*, clique em `Help->Markdown` no menu acima.

Claro, o Jupyter permite códigos e visualizações mais complexas como no exemplo abaixo. Execute a célula para ver o resultado.

```
import matplotlib.pyplot as plt
import numpy as np
from matplotlib.patches import Ellipse

%matplotlib inline

NUM = 250

ells = [Ellipse(xy=np.random.rand(2) * 10,
                width=np.random.rand(), height=np.random.rand(),
                angle=np.random.rand() * 360)
        for i in range(NUM)]

fig, ax = plt.subplots(subplot_kw={'aspect': 'equal'})
for e in ells:
    ax.add_artist(e)
    e.set_clip_box(ax.bbox)
    e.set_alpha(np.random.rand())
```


Documentação - Markdown

Formatação

É possível formatar a célula de documentação de diversas formas. Para ver como esta célula que você está lendo foi formatada, clique duas vezes para iniciar a edição. A linguagem usada para fazer a formatação é chamada **Markdown**. Ela permite diversas coisas como:

- Listas
- **Itálico**
- *****Negrito*****
- Fórmulas matemáticas: $f(x)=x^2+2x+\log_2 x$
- Links: [\[UTFPR\]\(http://www.utfpr.edu.br\)](http://www.utfpr.edu.br)

Para conhecer outras possibilidades de uso de **Markdown**, clique em ``Help->Markdown`` no menu acima.

Exercícios!

- Revise o conteúdo e faça os exercícios do notebook
01a-Jupyter Notebook_Introdução.ipynb
- Link:
<https://gitlab.com/luizcelso/datascience>
- Use o Binder clicando no link:



Exercícios!

- Execute o notebook anterior localmente:

```
git clone https://gitlab.com/luizcelso/datascience.git  
cd datascience/  
jupyter-notebook
```

Pacotes Python

- Pandas – operações de manipulação de dados estruturados
- NumPy – estruturas e operações numéricas
- Scikit Learn - machine learning
- Matplotlib – plotagem de gráficos
- Veremos também: StatsModels, NetworkX, NLTK, SeaBorn, Bokeh...

Instalação de Pacotes

- Recomendado: Use o Anaconda Navigator
- Pela linha de comando:
conda install <pacote>
ou
pip install <pacote>
- Instalar sem privilégios de administrador:
pip install --user <pacote>

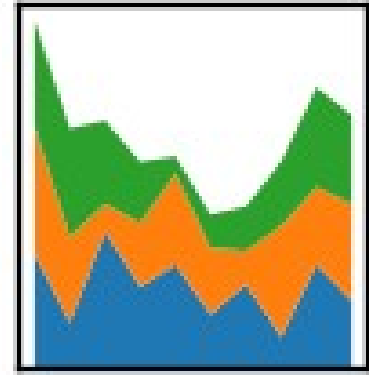
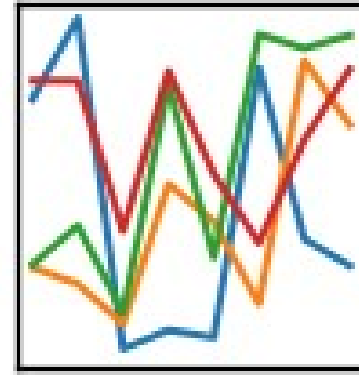
Instalação - Exemplo

- Aqui no Laboratório:

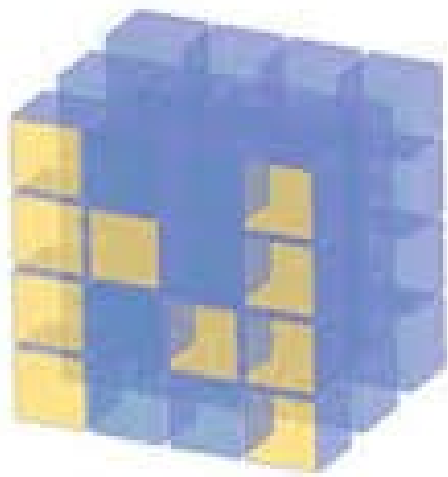
```
pip install --user pandas matplotlib
```

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



- Used for **structured data** operations and **manipulations**. It is extensively used for data munging and preparation.
- Have been instrumental in boosting Python's usage in data scientist community.

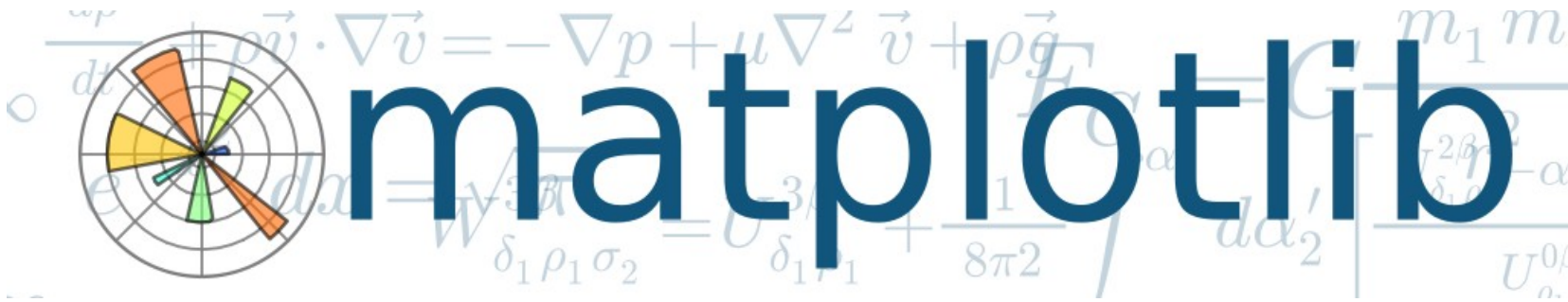


NumPy

- The most powerful feature of NumPy is n-dimensional array.
- This library also contains basic linear algebra functions, Fourier transforms, advanced random number capabilities and tools for integration with other low level languages like Fortran, C and C++
- Efficient processing



- Scikit-learn is a free software machine learning library for the Python programming language
- Features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN
- Designed to interoperate with NumPy and SciPy



- For plotting vast variety of graphs, starting from histograms to line plots to heat plots..
- You can use Pylab feature in Jupyter notebook (ipython notebook –pylab = inline) to use these plotting features inline.
- Alternatives: Bokeh, Seaborn



spyder

Data Science IDE

7)

Arquivo Editar Pesquisar Código Executar Depurar Consoles Projetos Ferramentas Ver Ajuda

C:\Users\Luiz

Editor - C:\Users\Luiz\.spyder\temp.py

```
1 # -*- coding: utf-8 -*-
2 """
3 Spyder Editor
4
5 Este é um arquivo de script temporário.
6 """
7 from pylab import *
8
9 x = np.linspace(0, 5, 10)
10 y = x ** 2
11
12 figure()
13 plot(x, y, 'r')
14 xlabel('x')
15 ylabel('y')
16 title('title aaaa')
17 show()
18
19 #%%
20
21 figure()
22 plot(y, x, 'r')
23 xlabel('x')
24 ylabel('y')
25 title('title xxxxxxx')
26 show()
27
28 #%%
29
```

Explorador de variáveis

Nome	Tipo	Tamanho	Valor
e	float	1	2.718281828...
euler_gamma	float	1	0.577215664...
pi	float	1	3.141592653...

Explorador de variáveis Explorador de arquivos Ajuda

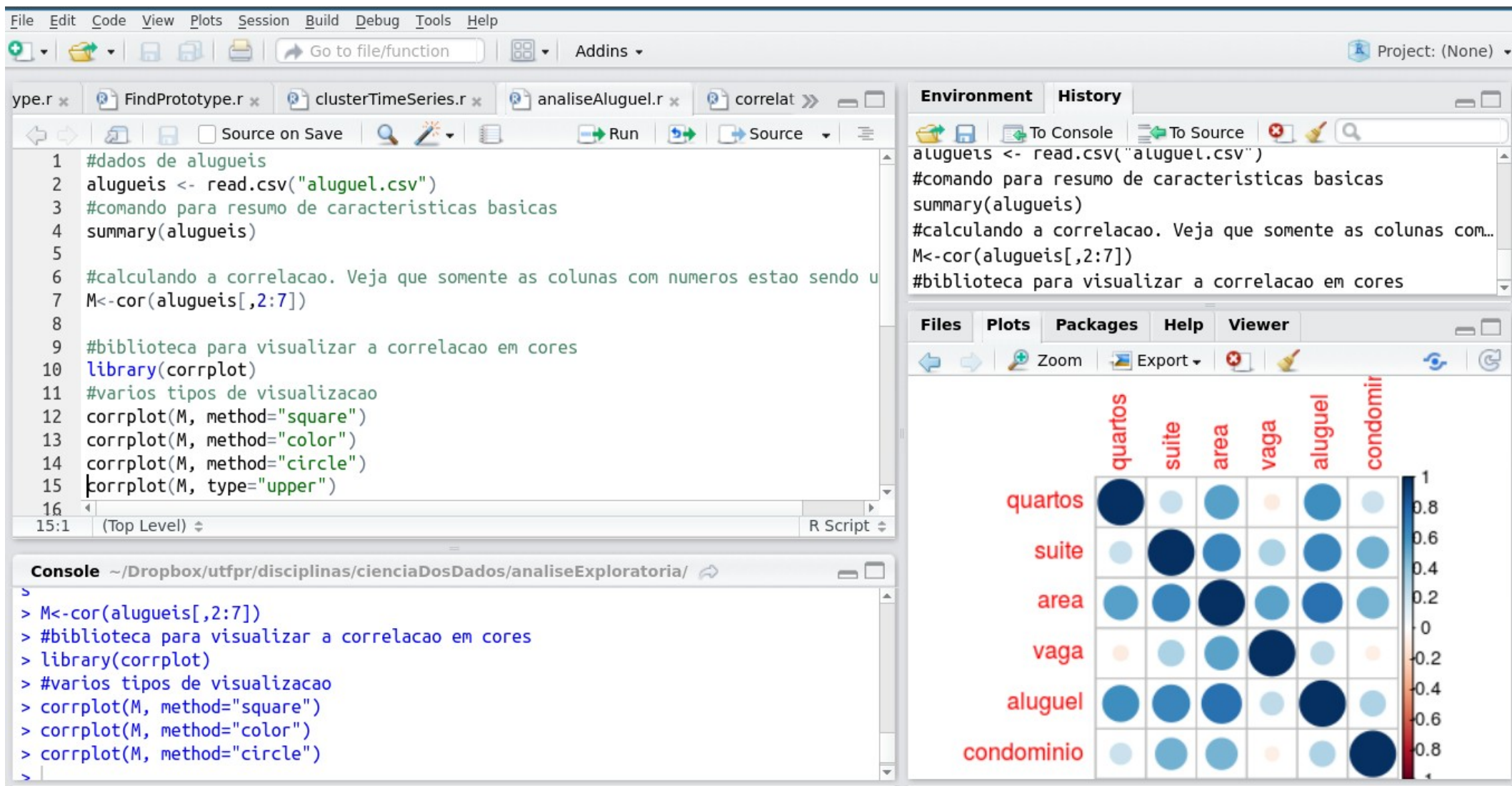
Console IPython

```
....: title('title xxxxxxx')
....: show()
....:
....:
```

Console Python Log do histórico Console IPython

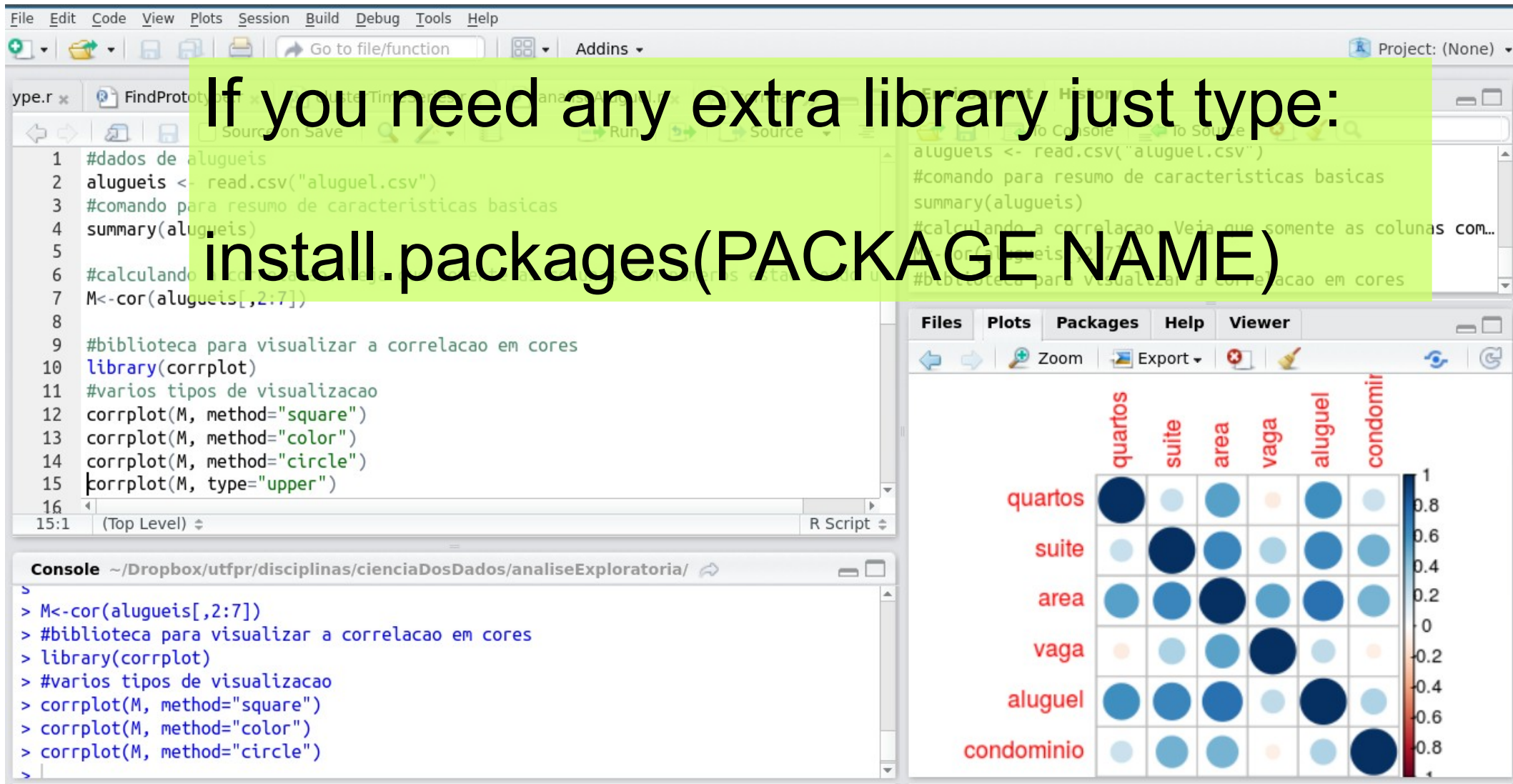
Permissões: RW Fim de linha: CRLF Codificação: UTF-8 Linha: 21 Coluna: 9 Memória: 79 %

R Studio



R Studio

If you need any extra library just type:
install.packages(PACKAGE NAME)



BI Tools

