



Ciencia de Datos  
Ecuador

# Introducción al análisis de datos con Python

Marcela Sofía  
Cevallos





# 1. ¿Qué es Python?

Python es uno de los lenguajes de programación más utilizados por los Analistas y Científicos de Datos. Nos permite realizar análisis exploratorios de datos, visualizaciones, modelos de machine learning, deep learning y mucho más.

Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma así como un lenguaje interpretado, dinámico y multiplataforma.

## **Un poco de historia:**

Python fue creado a finales de los ochenta por Guido van Rossum en el Centro para las Matemáticas y la Informática (CWI, Centrum Wiskunde & Informatica), en los Países Bajos, como un sucesor del lenguaje de programación ABC, capaz de manejar excepciones e interactuar con el sistema operativo Amoeba.

El nombre del lenguaje proviene de la afición de su creador por los humoristas británicos Monty Python.



## 1.1. Ventajas y “Desventajas” de Python

### Ventajas

- Simplificado y rápido
- Elegante y flexible
- Popular en las ramas de ciencia de datos, machine learning, e inteligencia artificial
- Ordenado y limpio, portable
- Comunidad activa
- Open source o de código abierto
- Multiplataforma
- Curva de aprendizaje

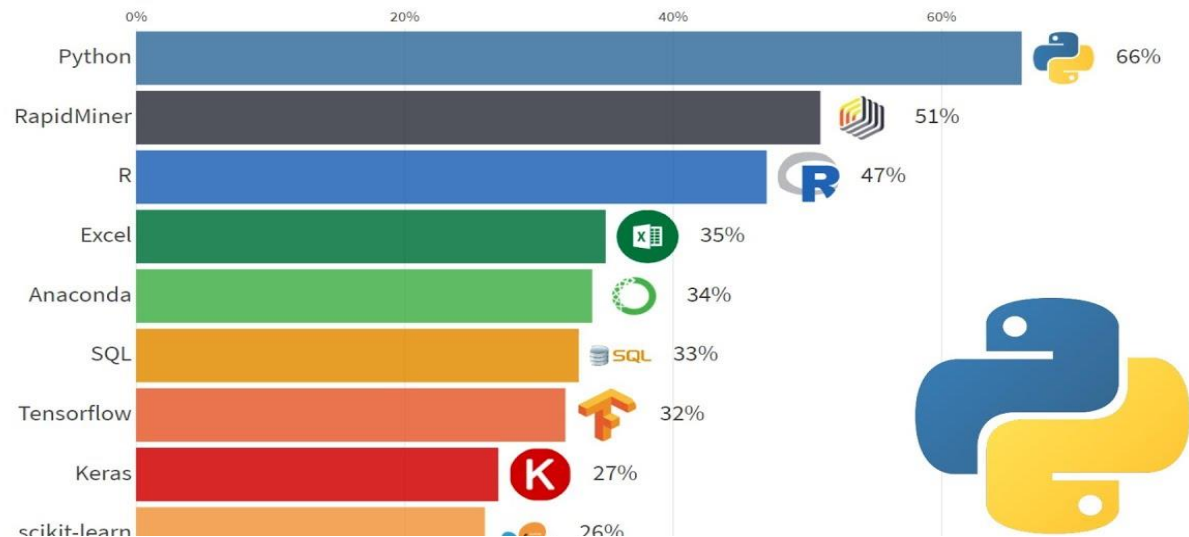
### “Desventajas”

- Hosting
- Lentitud al ejecutar múltiples hilos
- La curva de aprendizaje suele ser mayor a un software con interfaz de botones



## 1.2. ¿Por qué aprender Python?

Python es un lenguaje de programación interpretado gratuito cuya principal filosofía es que sea legible por cualquier persona con conocimientos básicos de programación. Además, posee una serie de características que lo hacen muy particular y que, sin duda, le aportan muchas ventajas y están en la raíz de su uso tan extendido:



Su página web se encuentra en el puerto: [www.python.org/](http://www.python.org/) donde existe toda la información y documentación necesaria, sin embargo se recomienda instalar Python por medio de la distribución de Anaconda



# 1.3 Anaconda

File Help


**ANACONDA.NAVIGATOR** Sign in

[Home](#)

[Environments](#)

[Learning](#)




[Community](#)

 [Join Now](#)









Discover premium data science content

[Documentation](#)

[Anaconda Blog](#)

Applications on base (root) Channels Refresh

 <b>CMD.exe Prompt</b> 0.1.1 Run a cmd.exe terminal with your current environment from Navigator activated <a href="#">Launch</a>	 <b>Datalore</b> Online Data Analysis Tool with smart coding assistance by JetBrains. Edit and run your Python notebooks in the cloud and share them with your team. <a href="#">Launch</a>	 <b>IBM Watson Studio Cloud</b> IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling. <a href="#">Launch</a>	 <b>JupyterLab</b> 3.0.14 An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture. <a href="#">Launch</a>
 <b>Jupyter Notebook</b> 6.4.0 Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis. <a href="#">Launch</a>	 <b>Powershell Prompt</b> 0.0.1 Run a Powershell terminal with your current environment from Navigator activated <a href="#">Launch</a>	 <b>Qt Console</b> 5.1.0 PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more. <a href="#">Launch</a>	 <b>Spyder</b> 5.0.3 Scientific PYTHON Development Environment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection Features <a href="#">Launch</a>

Windows taskbar: Escribe aquí para buscar | 17:14 2021/07/14



# 1.4. Jupyter Notebook

Jupyter Notebook (anteriormente IPython Notebooks) es un entorno informático interactivo basado en la web para crear documentos de Jupyter notebook.

El término "notebook" puede hacer referencia coloquialmente a muchas entidades diferentes, principalmente la aplicación web Jupyter, el servidor web Jupyter Python o el formato de documento Jupyter según el contexto.

Un documento de Jupyter Notebook es un documento JSON, que sigue un esquema versionado y que contiene una lista ordenada de celdas de entrada/salida que pueden contener código, texto (usando Markdown), matemáticas, gráficos y texto enriquecidos, generalmente terminado con la extensión ".ipynb".

The screenshot shows a Jupyter Notebook interface with the title "jupyter r\_\_Boston\_Crimes\_1591371215". The top bar includes a "Logout" button and a "Python 3" indicator. The menu bar contains "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help". The toolbar includes icons for file operations, a "Run" button, and a "Markdown" dropdown. The notebook content is titled "Crímenes de Boston" and includes a welcome message in Spanish. It shows two code cells: the first imports pandas, numpy, matplotlib, seaborn, and folium; the second loads a CSV file from a GitHub repository. The output of the second cell is a table of crime data.

### Crímenes de Boston

Bienvenido Agente Data, Big Data. Hemos sido contactados por el departamento de Policía de la ciudad de Boston. El alcalde cree que existe la posibilidad de predecir la cantidad de crímenes que suceden en su ciudad. Tu objetivo es analizar los datos de la ciudad y buscar patrones que nos permitan determinar la factibilidad de un modelo.

Esta misión es de suma importancia! Mucha suerte!!!

```
In [43]: import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
import folium
from folium.plugins import HeatMap

%matplotlib inline
```

```
In [44]: # carga La data
url = 'https://raw.githubusercontent.com/patofw/imf_master/master/Google_Colab/crimes.csv'
data = pd.read_csv(url, encoding='latin-1', sep = ';')

# cabecera
data.head()
```

Out[44]:

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	REPORTING_AREA	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOURL	UCR_PA
0	I182070304	1107	Fraud	905	1/11/2017 0:00	2017	11	Wednesday	0	Part 1
1	I182070115	3114	Investigate Property	793	11/10/2017 17:43	2017	10	Wednesday	17	Part Th



## 2. El Titanic dataset

### Un poco de historia:

El dataset del Titanic es uno de los más famosos dentro de Ciencia de Datos. Está basado en el hundimiento del Titanic en 1912.

La información en la base de datos contiene los nombres de los pasajeros, su edad, su sexo, si viajaban con sus hermanos, los puertos de embarque, y una columna que indica si el pasajero sobrevivió o no al hundimiento del Titanic.





## 2.1. La base de datos

La base de datos se encuentra en Kaggle en el siguiente link:

<https://www.kaggle.com/c/titanic>

Y es una muestra de los datos reales del hundimiento del Titanic.

A priori por historia sabemos que:

- No existían suficientes botes salvavidas a bordo
- Murieron 1502 de 2224 pasajeros y tripulación **(67.5%)**
- Se priorizó a un cierto grupo de pasajeros sobre otros

EDA

```
df.head(10) # El número en el paréntesis le indica al método cuántas observaciones desplegar
```

	id_pasajero	supervivencia	clase	nombre	sexo	edad	hermanos_esposos	padres_ninos	ticket	precio	cabina	puerto
0	1	0	3	Braund, Mr. Owen Harris	masculino	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	femenino	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	femenino	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	femenino	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	masculino	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	masculino	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	masculino	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	masculino	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	femenino	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	femenino	14.0	1	0	237736	30.0708	NaN	C





## 2.3. Objetivos y preguntas

1. ¿Cuál fue la tasa de supervivencia de las personas que viajaron en el Titanic? Concuerda con el dato histórico?
2. ¿Cuál fue la edad promedio de los pasajeros a bordo del Titanic?
3. ¿Cuál es la proporción de sobrevivientes por género a bordo del Titanic?
4. ¿Cuál fue la clase a bordo del Titanic con mayor probabilidad de supervivencia?
5. ¿Cuánto costaba en promedio un ticket a bordo del Titanic?
6. Bonus:
  1. Jack y Rose estaban a bordo del Titanic?

