

Clase 2: Lenguajes de programación para el análisis de datos

*Herramientas de programación para el análisis de
datos
2022*

¿Qué lenguajes de programación conocen?

<https://www.menti.com/dvoixvbyq9>

9110 8680



¿Qué se hace con estos lenguajes?
(En oraciones de una o dos palabras)

<https://www.menti.com/dvoixvbyq9>

9110 8680



Hoy

¿Cuáles son los principales LP usados para el análisis de datos?

¿Dónde suelen usarse cuáles lenguajes?

¿Cuáles son sus ventajas y desventajas?

¿Cuál es el mejor LP para cada necesidad?

¿Qué LP debo usar en mi trabajo?

¿Qué son los scripts?

¿Partir de cero?

¿Qué es lo común en mi disciplina?

¿Qué usa el equipo de trabajo?

¿Hay datos georreferenciados?

¿Qué volumen de datos?

¿Hay operaciones que se repiten con alguna frecuencia?

Libre

Pago

Economía



Estadística



STATA

IBM SPSS

python™

sas

Ingeniería



Negocios



Power BI



Visual Studio

+ a b l e a u

QGIS



Estudios urbanos y ambientales

Big

Scala

¿Qué quiero hacer con datos?

recolectar

explorar

limpiar

describir

relacionar

consultar

agregar - extraer

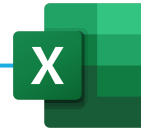
visualizar - presentar

clasificar - agrupar

ajustar

inferir

ML



¿Eficiencia?



+ a b l e a u



STATA



Algunas comparaciones entre los LP (1)

Código abierto

- *Python*
 - Programa multifuncional
 - Cuenta con un gran número de bibliotecas
 - Sintaxis simple
 - Fácil de integrar con otros programas
 - Amplias comunidades de apoyo
- *R*
 - Enfocado al análisis de datos y estadística
 - Bibliotecas de gráficos son las más usadas
 - Fácil de integrar con otros programas
 - Amplias comunidades de apoyo

Licencia comercial

- *Stata*
 - Es un entorno fácil gracias a su interfaz de usuario (GUI)
 - Los procedimientos estadísticos se encuentran casi que en su totalidad
 - Sus comandos están pensados para el análisis de datos
 - Amplias comunidades de apoyo
 - Comparativamente económico para sus competidores

Algunas comparaciones entre los LP (2)

Código abierto

- *Python*
 - Algunos procesos estadísticos no se han desarrollado lo suficiente (pero muy avanzados)
 - Existe una barrera de entrada
- *R*
 - Algunos paquetes no son tan estables como el programa en general
 - Altos requerimientos de hardware para procedimientos con bases de datos muy grandes
 - Existe una barrera de entrada

Licencia comercial

- *Stata*
 - Hasta antes de la versión 16 su integración con otros programas resulta engorrosa
 - Hasta antes de la versión 16 solo era posible abrir una base de datos por sesión
 - Puede ser lento para incorporar nuevos procedimientos estadísticos

¿Cuáles son los LP más usados?

Tamaño de las comunidades

		Most popular in	Least popular in
Javascript*	12.4M	Web, Cloud	DS/ML, AR/VR
Python	9.0M	DS/ML, IoT apps	Mobile, Web
Java	8.2M	Mobile, Cloud	DS/ML, Web
C/C++	6.3M	IoT apps, IoT devices, AR/VR	Web, Cloud, Mobile
PHP	6.1M	Web, Cloud	DS/ML, Mobile
C#	6.0M	Games, AR/VR, Desktop	DS/ML, Mobile
Visual development tools	2.8M	Desktop, AR/VR	Cloud, Web
Swift	2.4M	Mobile, AR/VR	Cloud, IoT devices
Kotlin	2.3M	Mobile, AR/VR	DS/ML, Desktop
Go	1.5M	Cloud, AR/VR	DS/ML, Web
Ruby	1.5M	IoT apps, Cloud	DS/ML, apps for 3rd party ecosystems
Objective C	1.4M	AR/VR, Mobile	Desktop, IoT devices
Rust	0.8M	AR/VR, IoT apps	Web, Cloud
Lua	0.8M	AR/VR, Games	Mobile, apps for 3rd party ecosystems

Demanda de empleadores

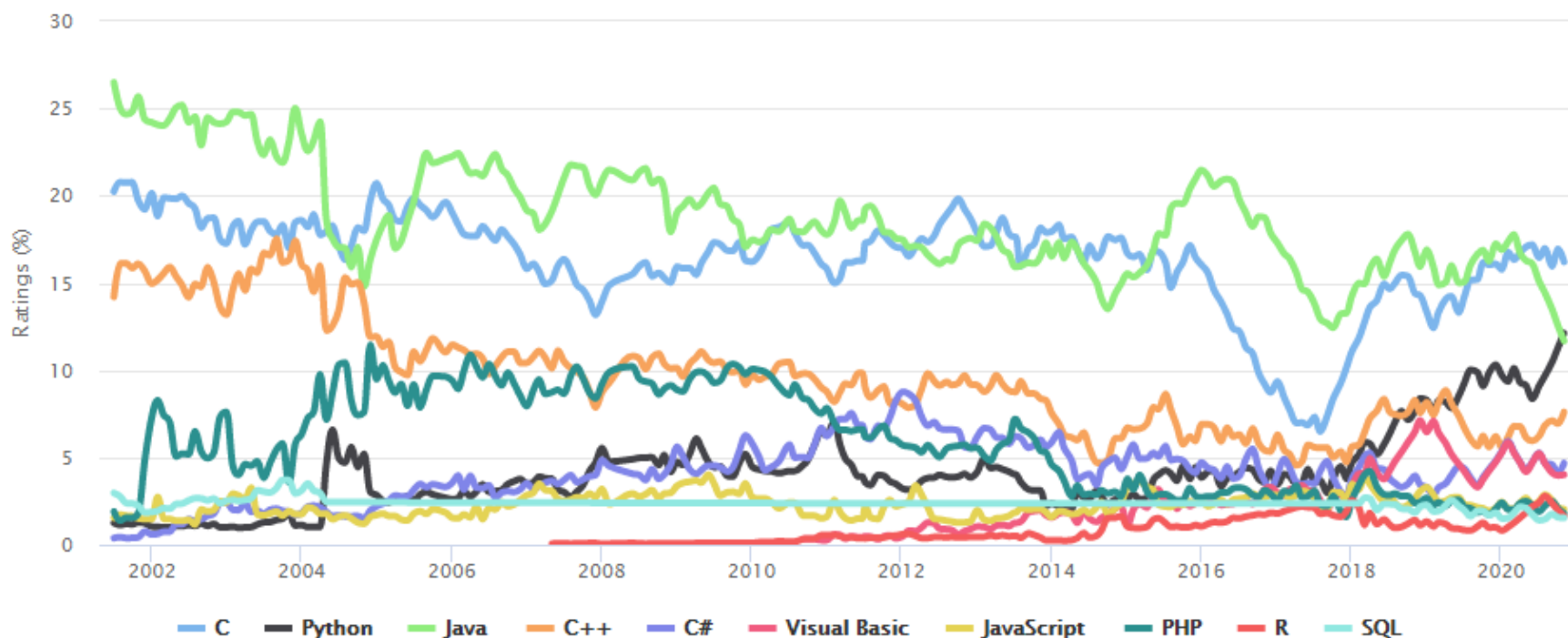
Rank	Language	Score
1	Python▼	100.0
2	C▼	98.0
3	Java▼	97.1
4	Go▼	87.2
5	C++▼	85.2
6	R▼	80.4
7	Swift▼	70.1
8	SQL▼	69.4
9	Ruby▼	67.4
10	Matlab▼	64.6

Fuente: Developer Economics – Q3/2020

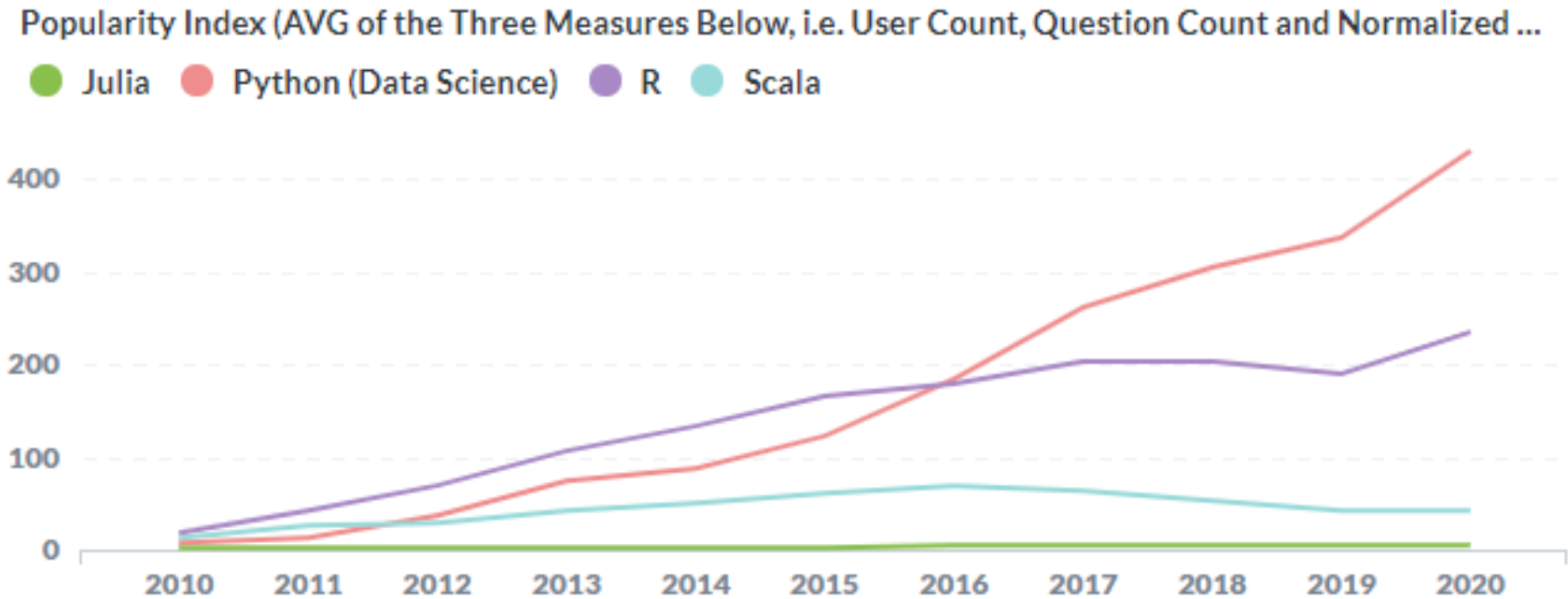
Popularidad de lenguajes

TIOBE Programming Community Index

Source: www.tiobe.com

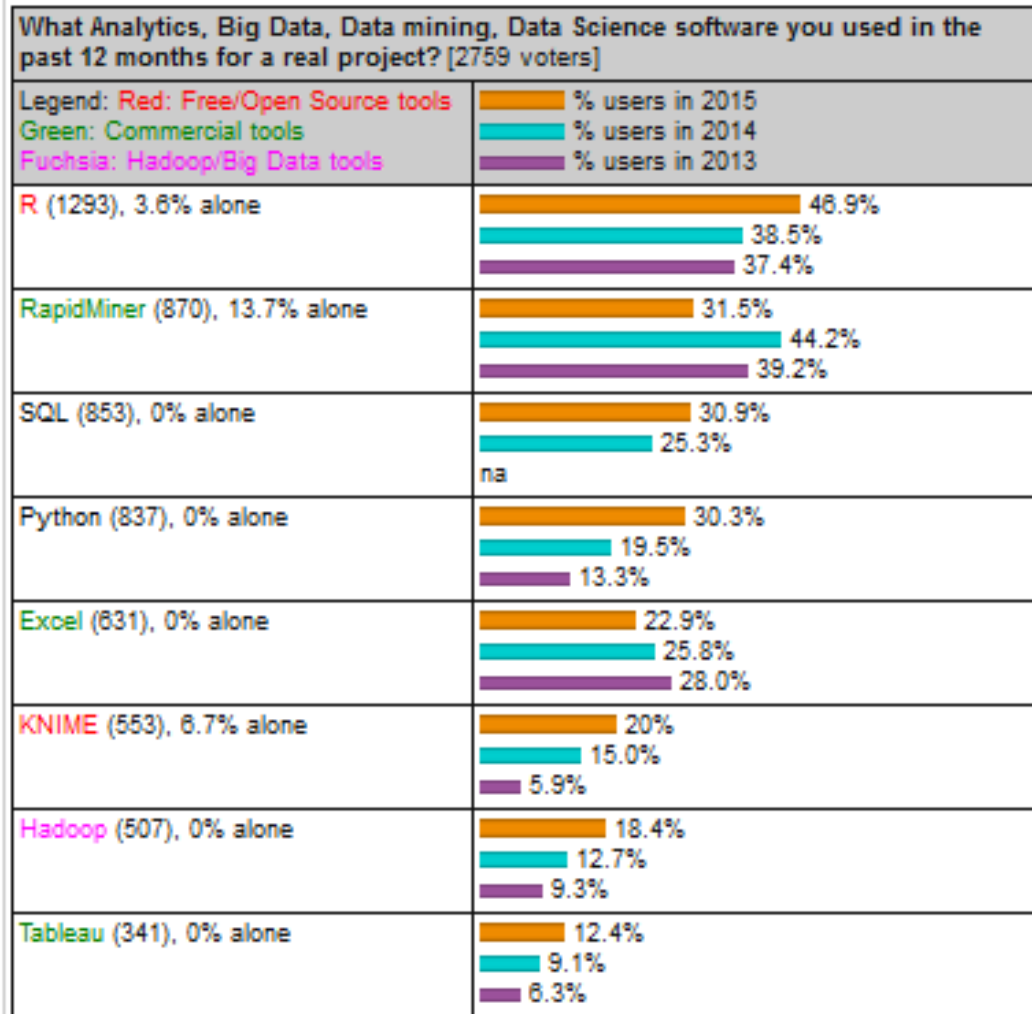


Algunos usados en datos



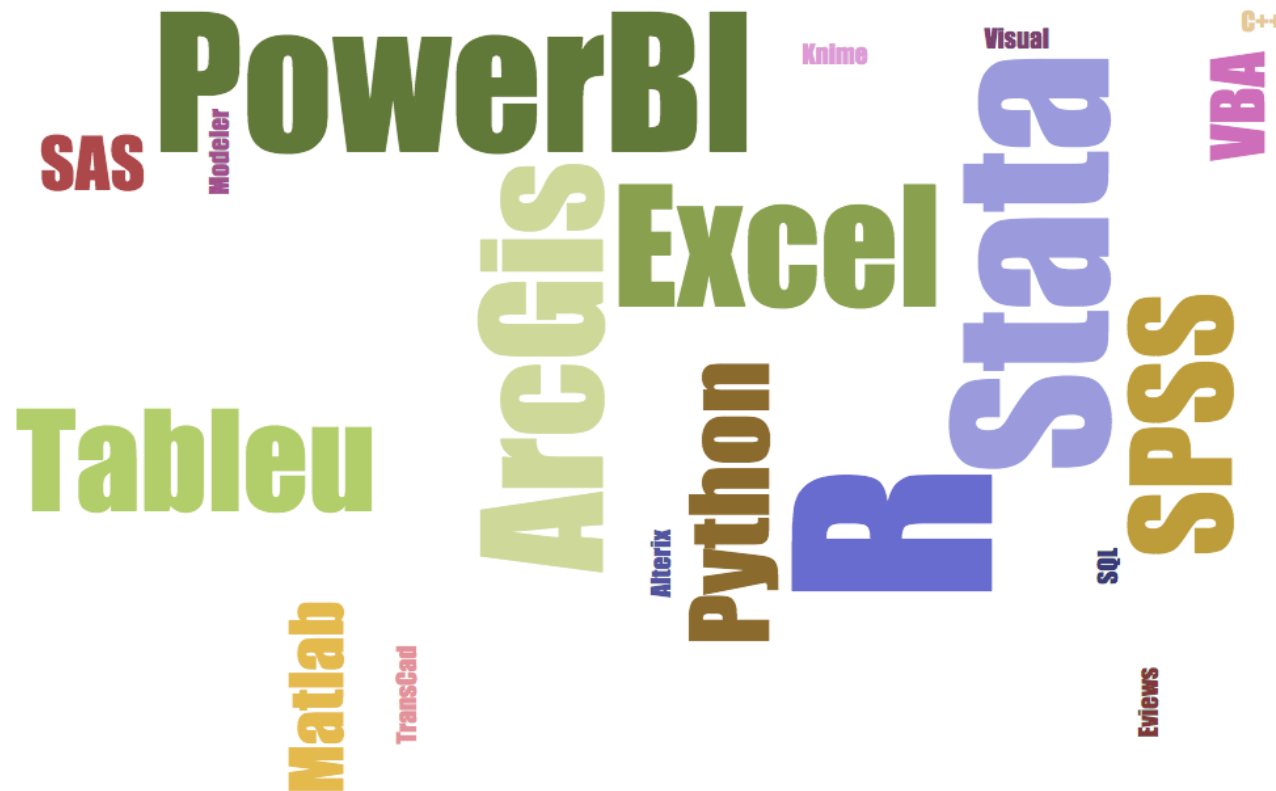
Fuente: Stack Overflow

Algunos usados en datos



Fuente: r4stats

¿Qué se usa en Colombia?



Fuente: los resultados corresponden a un corto y no representativo sondeo

SECTOR PRIVADO



SECTOR PÚBLICO

¿Cuál es el mejor LP?

- Es difícil establecer si existe un LP que sea superior a los demás debido a que cada uno tiene sus ventajas y sus debilidades.
- En este sentido, de acuerdo a las necesidades de la tarea, un programa estadístico será más eficiente que otro.
- Por ejemplo, si necesitamos calcular el promedio de dos números, usar Excel es una herramienta poderosa.

¿Cuál es el mejor LP?

- Se debe tener en cuenta que las personas se terminan especializando en un lenguaje debido a la curva de aprendizaje.
- Así, a menos que sea un procedimiento que el programa sea incapaz de realizar, deberíamos potenciar nuestra experiencia en un mismo lenguaje.
- Sin que lo anterior signifique que se deje estar al tanto de nuevos desarrollos de los otros programas.

¿Cuál es el mejor LP?

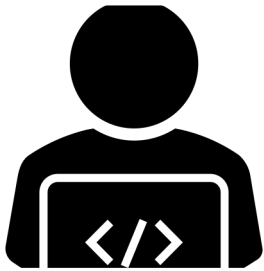
- En términos generales, la complejidad de la tarea define el programa que debería usar.
- Para un análisis de datos convencional es eficiente realizarlo en R o Stata porque sus finalidades son precisamente estas.
- Por otro lado, tareas que impliquen procedimientos más novedosos como la minería de texto, quizá sea más práctico emplear Python o R.
- Para rutinas de aprendizaje de máquinas Python o R son una gran solución.

¿Qué son los *scripts*?

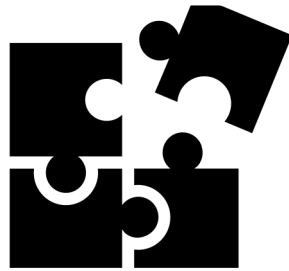
- Los diferentes programas cuentan con ventanas de comando en las cuales es posible introducir las órdenes.
- Esta práctica impide la fácil replicación de un conjunto de tareas. Por esta razón, es posible crear un archivo de texto que permita almacenar de manera secuencial las órdenes que necesitamos ejecutar para cumplir con nuestra tarea.
- Por ello existen los scripts

	Stata	R	Python
Nombre	Do file	Script	Script
Extensión	.do	.r	.ipynb

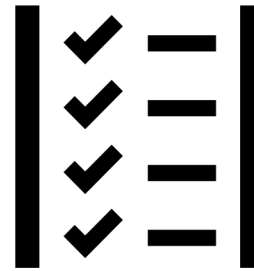
¿Qué son los *scripts*?



Incluir información sobre la autoría del código



Dividir en secciones



Comentar los procedimientos



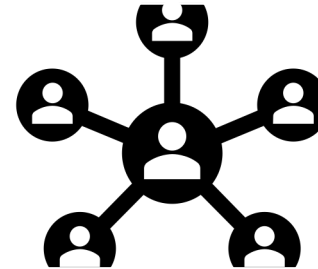
Ser cuidadoso con la tabulación del código



Llevar un registro de las modificaciones



Cada script debe ser ejecutada sin interrupción



Si existen grandes tareas no es recomendable incluir todo en el mismo archivo.

```

1  * Por: Miguel Andrés Garzón Ramírez
2  * 2 de febrero de 2021
3  * Versión 1
4
5  clear all
6  global dir "C:\Users\magse\Dropbox\Ejemplo"
7
8  ** Cargar datos
9  use "$dir\GEIH_migración_septiembre_2020_putexcel.dta", clear
10
11
12 ** Método actuando sobre la base de datos
13 putexcel set "$dir\results.xlsx", sh(collapse, replace) modify // como no se declara una carpeta de referencia se debe utilizar la macro
14 (global) con la dirección completa donde se va a guardar el archivo que estamos declarando aquí. El archivo se crea cuando en otro putexcel
15 se introduzca algún dato.
16 *set trace on
17 local r=1
18 putexcel A1=("area3") B1=("Edad") C1=("Freq.")
19 levelsof area3, local(area3_cat) // crea una lista con los valores de la variable area3 para iterar sobre ellos
20 foreach area3 of local area3_cat{
21     putexcel A`r'+1'=(`area3') // colocar el indicador de la categoría en la tabla
22     preserve
23     gen freq=1 // variable temporal para hacer el conteo de frecuencias
24     collapse (count) freq [i=Fex_c_2011] if area3 == `area3' & P755S3==3 & P756S3==3, by(edad_rank) // se detiene si el filtro no
25     tiene observaciones, es mejor iterar sobre las categorías de la variable area3. Crea una tabla de frecuencias similar lo que hace tab. Es
26     más sencillo con el comando contract, pero le tengo mas confianza a collapse
27     levelsof edad_rank, local(edad_rank_cat) // crear lista de edades que quedan en la tabla para colocar los datos fila a fila en excel.
28     local i=1 // indicador de la fila en la tabla creada con collapse
29     foreach cat of local edad_rank_cat{
30         local et_edad_rank: label edad_rank `cat' // tomar la etiqueta de valor del rango de edad utilizado en esta iteración
31         putexcel B`r'+1'=""et_edad_rank"" C`r'+1'=freq[`i'] // colocar en excel la etiqueta y el valor
32         local ++r // cambiar el indicador de fila de la tabla principal
33         local ++i // cambiar el indicador de fila de la tabla de cada valor de area3
34     }
35     restore
36 }
37 *

```