Motivación del uso de R

UCO ONLINE

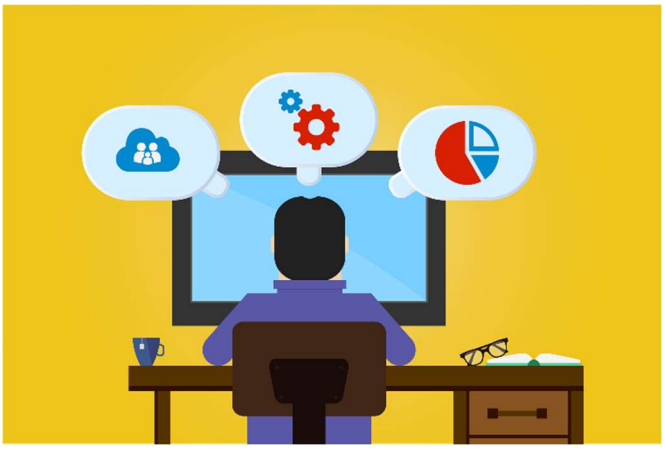
Motivación del uso de R

Máster en Ciencia de Datos

El presente curso pretende introducirnos al lenguaje de programación R, estando dirigido a estudiantes que nunca han usado R o ningún otro lenguaje de programación, e incluso no tienen conocimientos previos de probabilidad y estadística. Este curso tiene como propósito que adquieras los fundamentos del uso de R como un lenguaje de programación, desde sus conceptos más elementales hasta la iniciación en los paquetes más vinculados a ciencia de datos.

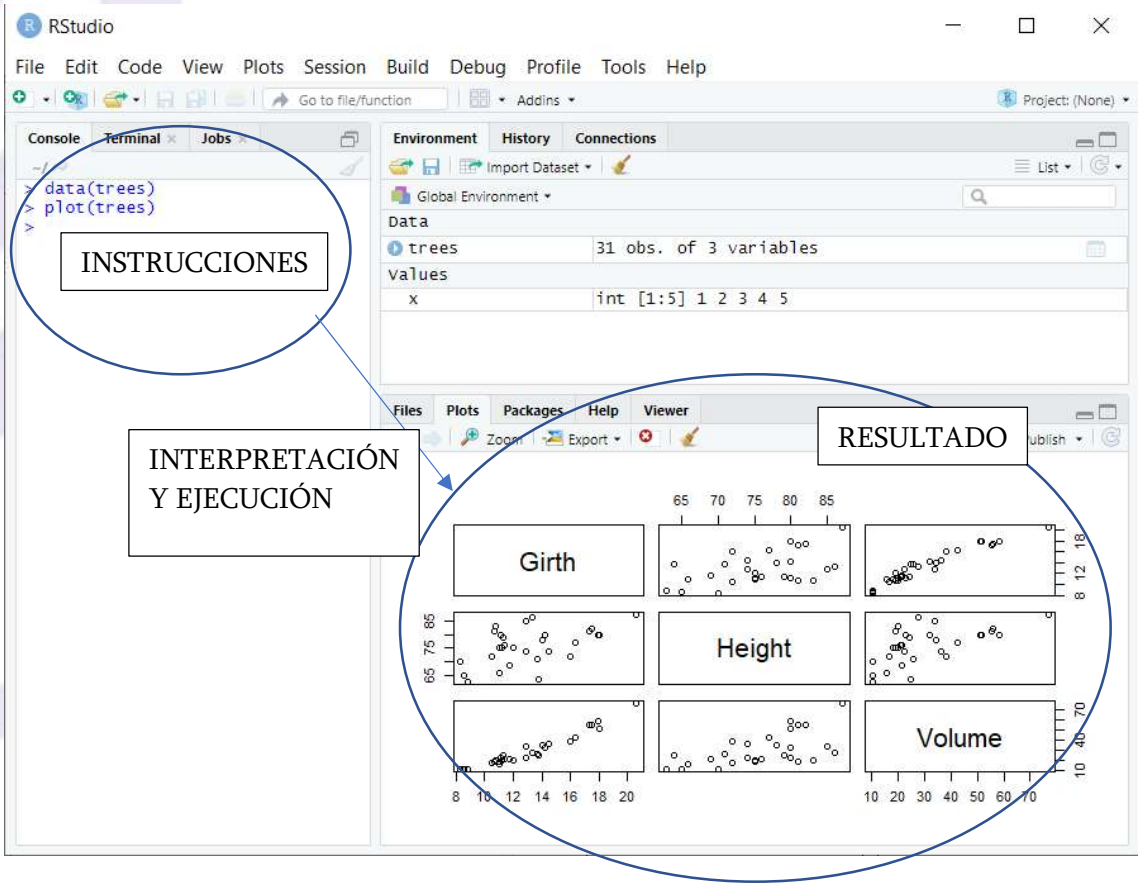
¿Qué es R?

R es un lenguaje de programación dedicado a la Estadística. Decimos que es un lenguaje de programación porque nos permite dar instrucciones, usando código, a nuestros equipos de cómputo para que realicen tareas específicas. Para ello sólo necesitaremos de un intérprete para este código: donde se compruebe que las instrucciones son correctas, sean ejecutadas estas instrucciones y se muestre el resultado de tal ejecución. A esto último es a lo que llamamos un entorno computacional o un entorno de desarrollo. En nuestro caso esta labor la cumplirá RStudio.



Proceso de trabajo con R

Para que podamos hacer algo en ese entorno necesitamos conocer la manera de escribir instrucciones que el software pueda interpretar y ejecutar. Eso es lo que aprenderemos a hacer en este curso. R es diferente a otros lenguajes de programación que por lo general están diseñados para realizar muchas tareas diferentes; esto es porque fue creado con el único propósito de hacer estadística.



Orígenes de R

R tiene sus orígenes en S, un lenguaje de programación creado en los Laboratorios Bell de Estados Unidos: los mismos laboratorios que inventaron el transistor, el láser, el sistema operativo Unix y algunas otras cosas más. Dado que S y sus estándares son propiedad de los Laboratorios Bell, lo cual restringe su uso, Ross Ihaka y Robert Gentleman, de la Universidad de Auckland en Nueva Zelanda, decidieron crear una implementación abierta y gratuita de S. Este trabajo, que culminaría en la creación de R

inició en 1992, teniendo una versión inicial del lenguaje en 1995 y en el 2000 una versión final estable.

¿Quién crea y mantiene R?

En el presente, el mantenimiento y desarrollo de R es realizado por el R Development Core Team, un equipo de especialistas en ciencias computacionales y estadística provenientes de diferentes instituciones y lugares alrededor del mundo:

https://www.r-project.org/contributors.html

La versión de R mantenida por este equipo es conocida como "base" y como su nombre indica, es sobre aquella que se crean otras implementaciones de R así como los paquetes que expanden su funcionalidad. Para lograr que R sea usado sin restricciones es distribuido de manera gratuita, a través de la Licencia Pública General de GNU, por lo que es software libre y de código abierto

https://www.r-project.org/about.html

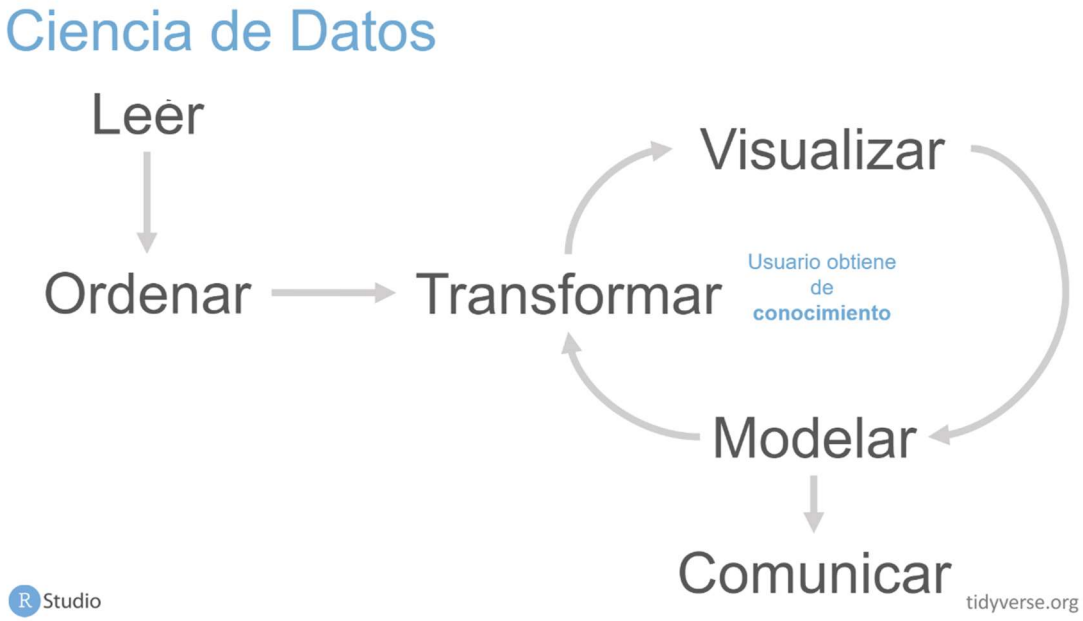
En la actualidad, el desarrollo de este lenguaje de programación se mantiene activo. La versión más reciente de R al momento de escribir este documento es la 4.0.4 que fue publicada en febrero de 2021 y diariamente son publicados nuevos paquetes y sus respectivas actualizaciones.

Aplicaciones de R: ciencia de datos.

R puede ser aplicado en múltiples ámbitos con resultados prometedores:

a) Como simple calculadora.
b) Estadística descriptiva.
c) Inferencia estadística.
d) Ajuste de modelos.
e) Estadística espacial.
f) Epidemiología.
g) Análisis de señales.
h) Genética.
i) Econometría.
j) Series temporales.
k) Ecuaciones estructurales.
l) Métodos bayesianos…

La ciencia de datos es un subconjunto de la Inteligencia Artificial que aborda principalmente las áreas interconectadas de estadística, métodos científicos y análisis de datos, todas las cuales se utilizan para extraer significado y conocimiento de los datos. El progreso tecnológico ha permitido la creación y almacenamiento de cantidades cada vez mayores de información, habiéndose incrementado notablemente los volúmenes de datos (imágenes, vídeos, documentos, etc.). R nos va a ayudar a revelar tendencias y generar conocimiento de esas grandes cantidades de datos que se podrá utilizar para tomar mejores decisiones acerca de lo que significan, revelan y reflejan esos almacenes de datos.



Propiedades de R:

R es un entorno y lenguaje de programación, distribuido bajo la licencia GNU GPL, con las siguientes características:

- *Entorno de trabajo:* es sobre todo un entorno de trabajo en el que se pueden manipular múltiples elementos: archivos de datos, gráficos, enlaces a recursos web, etc. Este entorno de trabajo cuenta con la indiscutible ventaja de ser idéntico en los tres grandes sistemas operativos: Linux, Windows y Mac.
- *Interfaz:* actúa como interfaz para procedimientos computacionales muy diversos. Es interactivo. Es una parte integral del proceso de investigación reproducible.
- *Interpretado y no compilado:* no tenemos que compilar nuestro código, sino que el intérprete de R lo ejecuta directamente (a diferencia de otros lenguajes de programación).
- *Basado en memoria:* R mantiene todos los objetos que definimos en nuestro programa en memoria.
- *Orientado a objetos:* R nos permite modelar conceptos del mundo real relevantes a nuestro problema, representándolos como clases y objetos que podemos hacer que interactúen entre sí. Todo en R es un "objeto".
- *Modelo de programación funcional:* Las funciones (conjunto de instrucciones agrupadas para realizar una tarea) en R se pueden manipular de forma sencilla.
- *Modular:* Es modular, construido a partir de mútiples "piezas" ajustadas a un formato estándar.
- *Extensible:* se puede extender a partir de la definición de funciones propias (conjunto de instrucciones agrupadas y reutilizables con una determinada funcionalidad definida), aparte de las de los numerosos paquetes que ya hay elaborados para R.
- *Colaborativo:* Puesto que R pertenece a un proyecto colaborativo y abierto, los propios usuarios pueden publicar paquetes (conjuntos de funciones agrupadas para un mismo campo o cometido) que extienden su configuración básica. Es software libre, abierto a la participación de quien desee aportar su conocimiento y experiencia.
- *Con capacidad de importación:* ofrece múltiples posibilidades para importar datos almacenados en distintos tipos de bases de datos. También presenta múltiples paquetes que permiten a R interactuar con otros lenguajes e intercambiar objetos con ellos.

¿Quién utiliza R?

R es un lenguaje relativamente joven pero que ha experimentado un crecimiento acelerado en su adopción durante los últimos quince años.

De acuerdo al TIOBE *programming community index* (2021), que es uno de los índices de más prestigio en el mundo en relación a la popularidad en el uso de lenguajes de programación, R es el lenguaje número 11 en popularidad, después de haber sido el lenguaje número 13 en el 2020. Esto es notable si consideramos que R es un lenguaje dedicado únicamente a la estadística, mientras que lenguajes como Python (número 3 en 2021) o C (número 1 en 2021) son lenguajes que pueden ser usados para todo tipo de tareas.

https://www.tiobe.com/tiobe-index/

La adopción de R se debe en gran medida a que permite responder preguntas mediante el uso de datos de forma efectiva, y como es un lenguaje abierto y gratuito, se facilita compartir código, crear herramientas para solucionar problemas comunes y que todo tipo de personas interesadas en análisis estadísticos puedan participar y contribuir al desarrollo y uso de R.

Por citar un ejemplo, es usado por Facebook para analizar la manera en que sus usuarios interactúan con sus muros de publicaciones para así determinar qué contenido mostrarles. Esta es una tarea muy importante en Facebook, pues las interacciones de los usuarios con publicidad y contenido pagado son la principal fuente de ingreso de esta compañía. Además de que su división de recursos humanos emplea esta herramienta para estudiar las interacciones entre sus trabajadores.

Google usa R para analizar la efectividad de las campañas de publicidad implementadas en sus servicios, por ejemplo, los anuncios pagados que te aparecen cuando buscas algo. Nuevamente, esta es una de las principales fuentes de ingresos de esta compañía. Así mismo, R también es usado para hacer predicciones económicas y otras actividades.

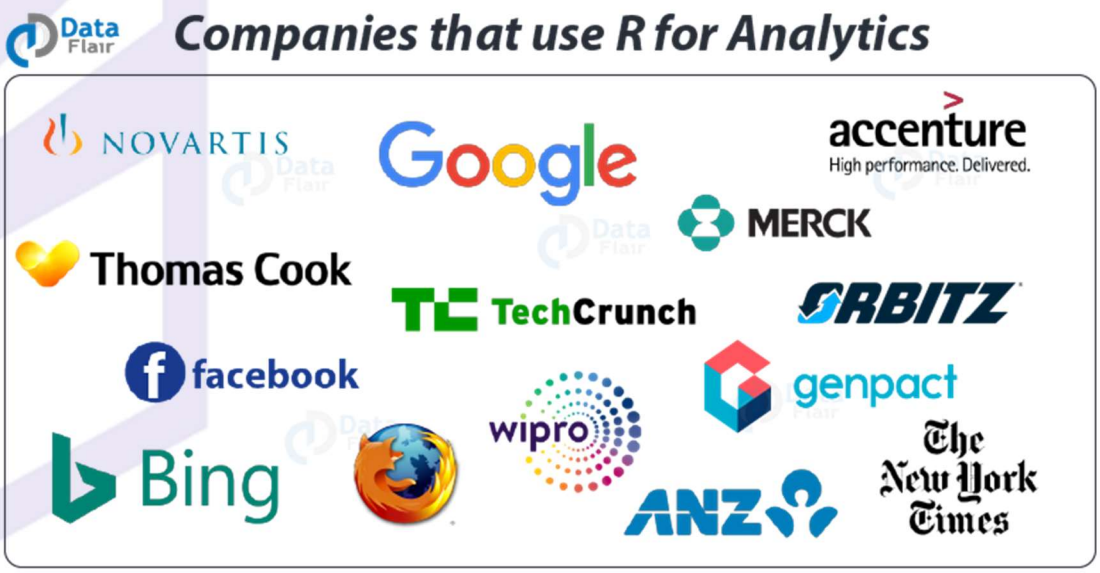
Microsoft ha adoptado por completo el lenguaje de programación R como herramienta de primera clase para científicos de datos. Microsoft adquirió y ahora desarrolla una versión propia de R llamada R Open, que ha hecho disponible para uso general del público. R Open es empleada para realizar todo tipo de análisis estadísticos, por ejemplo, para emparejar a jugadores en la plataforma de videojuegos Xbox.

https://mran.microsoft.com/open

https://docs.microsoft.com/es-es/azure/architecture/data-guide/technology-choices/r-developers-guide

Otras compañías que usan R de modo cotidiano son: Airbnb, Booking, Amazon, BBC, Mozilla, Netflix, etc.

https://github.com/ThinkR-open/companies-using-r



https://data-flair.training/blogs/r-applications/

Lo anterior ilustra algunas de las aplicaciones específicas de este lenguaje y de manera general podemos decir que R es usado para procesar, analizar, modelar y comunicar y entrelazar datos entre sí.

Aunque R está diseñado para análisis estadístico, con el paso del tiempo los usuarios de este lenguaje han creado extensiones a R, llamadas paquetes, que han ampliado su funcionalidad. En la actualidad es posible realizar en R minería de textos, procesamiento de imagen, visualizaciones interactivas de datos y procesamiento de *Big Data*, entre muchas otras posibilidades.