PROBABILIDAD Y ESTADÍSTICA

Curso de R

Gustavo A. Colmenares
gcolmenares@yachaytech.edu.ec
gcolmena@gmail.com

1- Introducción

UNIVERSIDAD YACHAY TECH

SCHOOL OF MATHEMATICAL AND COMPUTATIONAL SCIENCES

Contenido y programación del Taller

1. Introducción.
2. Conociendo el entorno de trabajo.
3. Fundamentos del lenguaje.
4. Estructuras de datos.
5. Paquetes.
6. Crear un Proyecto.
7. Importar, exportar datos.
8. Factor.
9. Gráficos con R Base.
10. EDA (Exploratory Data Analysis)

Modalidad: Virtual.

Fecha de inicio: 04/04/2022

Fecha de finalización: 08/04/2022

Horario: 9:00 – 12:00

Tiempo en contacto con capacitador: 15 horas.

Tiempo de trabajo autónomo: 15 horas

Tipo y horas de certificación: Por aprobación.

UNIVERSIDAD YACHAY TECH

SCHOOL OF MATHEMATICAL AND COMPUTATIONAL SCIENCES

Gustavo A. Colmenares gcolmenares@yachaytech.edu.ec / gcolmena@gmail.com

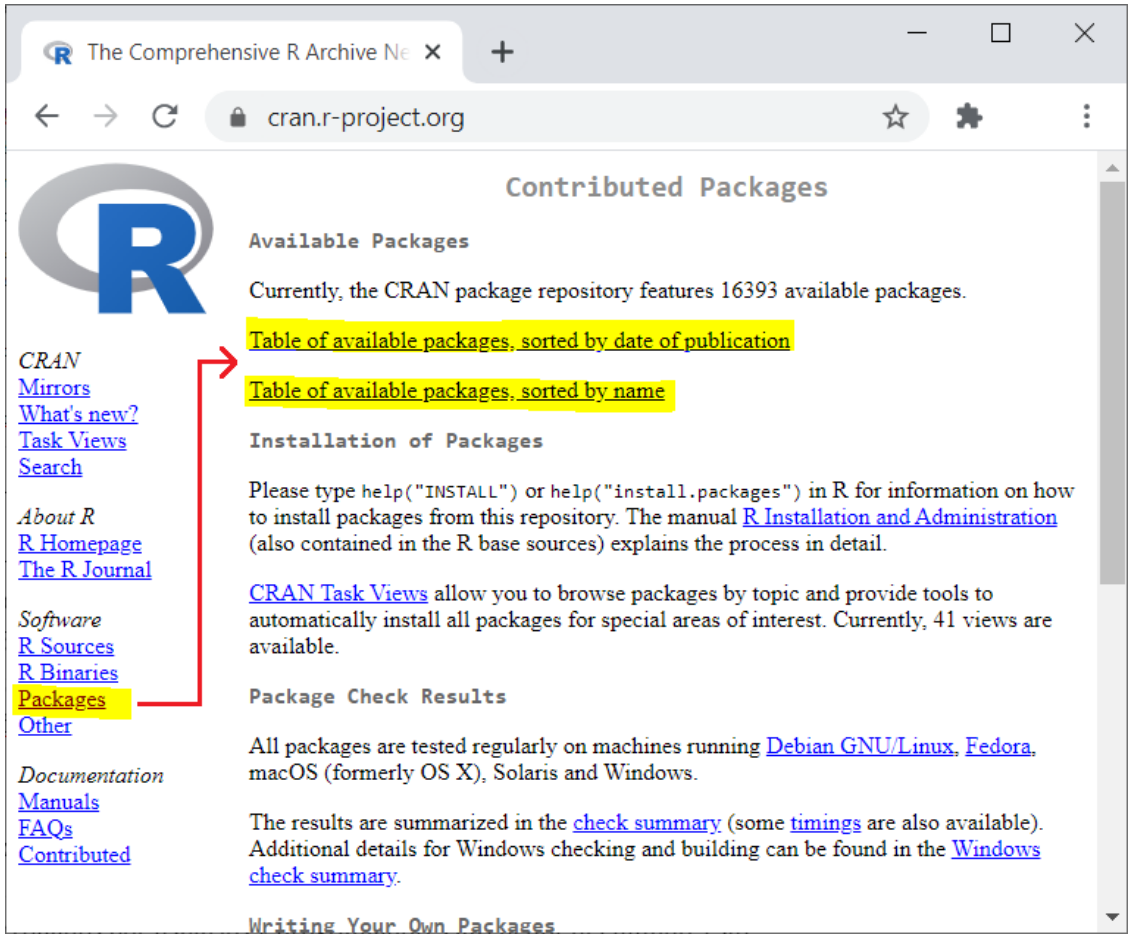
¿Qué es R?

- **R** es un entorno de Open Source (licencia GNU GLP), lenguaje de programación orientado a objetos e interpretado, que es usado como un ambiente de programación en el que se aplican técnicas estadísticas.
- El término **entorno** en **R**, se refiere a un sistema totalmente planificado y coherente para el análisis de datos.

- **Open Source** quiere decir que cualquier usuario puede descargar y crear su código de manera gratuita, sin restricciones de uso.
- **Lenguaje de programación orientado a objetos** significa que las variables, datos, funciones, resultados, etc., se guardan en la memoria activa del computador en forma de objetos con un nombre específico. Esta característica permite aplicar cálculos a un conjunto de valores a la vez, sin la necesidad de utilizar un algoritmo más sofisticado como por ejemplo, una función bucle.
- **R** es un **lenguaje interpretado** (como Python) y no compilado (como Fortran o Pascal), quiere decir que los comandos escritos en el teclado son ejecutados directamente sin necesidad de construir un ejecutable.

¿Qué es R?

- Su desarrollo actual es responsabilidad del **R Development Core Team**. Forma parte del proyecto colaborativo y abierto **CRAN R Proyect** (por sus siglas en inglés Comprehensive R Archive Network) http://cran.r-project.org/ donde los usuarios pueden publicar paquetes que extienden su configuración básica (repositorio oficial de paquetes).
- El sitio de Internet CRAN además centraliza una gran cantidad de información del proyecto que incluye manuales de uso en varios idiomas, información sobre paquetes, grupos de desarrollo a nivel mundial y los archivos binarios pre-compilados de **R**, que pueden ser descargados para Windows, Linux (Debian, Mandrake, RedHat, SuSe) y Apple Mac.



Breve historia

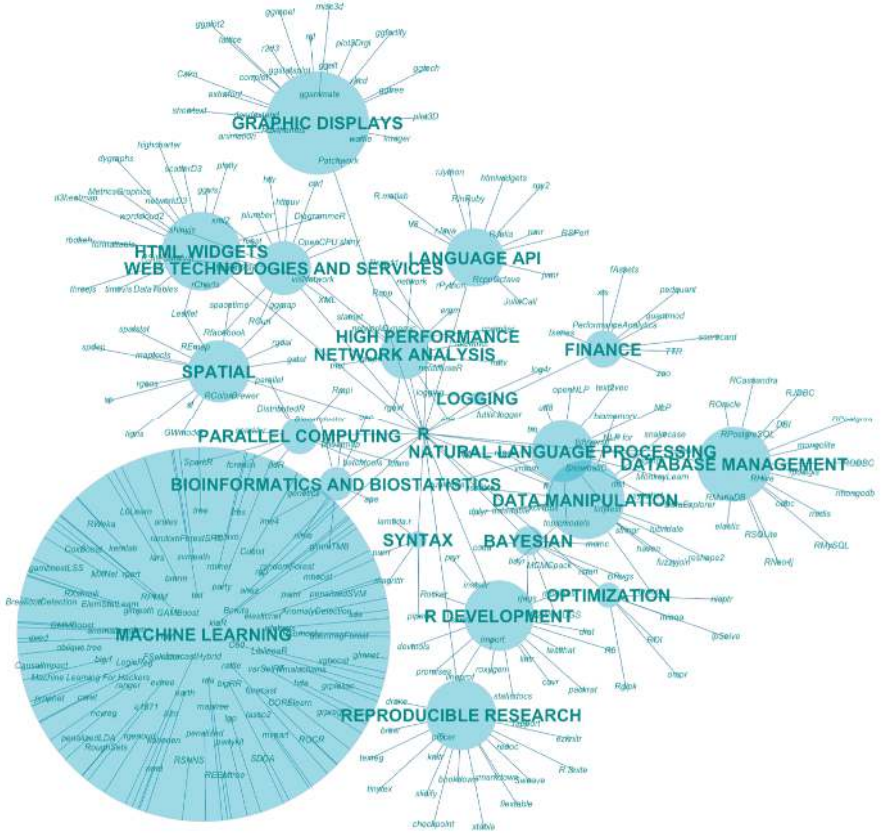
- Fue desarrollado inicialmente por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland (Nueva Zelandia) en 1992. Sin embargo, sus bases iniciales se remontan a un sistema para el análisis de datos desarrollado por John Chambers, Rick Becker, y otros colaboradores desde finales de 1970 en los Bell Laboratories de AT&T en Nueva Jersey, que denominaron el **lenguaje S**.
- Gentleman y Ihaka, combinaron las fortalezas del lenguaje S y Scheme (es un dialecto minimalista de la familia de lenguajes de programación Lisp) generando una implementación moderna a la cual denominaron simplemente como **R**. El nombre viene de la primera inicial de los nombres de los dos creadores (¿mito?).
- Luego de la creación de **R** (en 1992), se da un primer anuncio al público del **software R** en 1993.
- En 1995 Martin Mächler, de la Escuela Politécnica Federal de Zúrich, convence a Ross y Robert a usar la Licencia GNU para hacer de **R** un software libre. Paralelamente se inicia la comunidad pública de desarrolladores de **R**.
- En febrero del año 2000, luego de considerar al software completo y lo suficientemente estable, se libera la versión 1.0.

¿Cuáles son las ventajas de R?

- Fue **desarrollado por estadísticos.** Por lo tanto posee una gran colección de herramientas para el análisis, manipulación, almacenamiento y visualización de datos de forma precisa y productiva.
- Es **gratuito**. Esto significa que cualquiera puede instalarlo en cualquier organización sin necesidad de comprar una licencia. Por lo tanto no necesita buscar versiones piratas.
- Es de **código abierto**, lo que implica que los usuarios tienen la libertad de ejecutar, copiar, distribuir, estudiar, modificar y mejorar el software.
- Es **multiplataforma** (Windows, Linux, Mac). Incluso existen versiones de R para High Performance Computing (HPC) y en general ambientes multiprocesador (paralelismo).
- Es uno de los paquetes estadísticos de **mayor crecimiento** respecto de su uso en diferentes disciplinas.
- Posee uno de los **mejores sistemas para graficar** existentes en la actualidad.
- Puedes crear **aplicaciones web interactivas** (apps) con la herramienta **Shiny**.

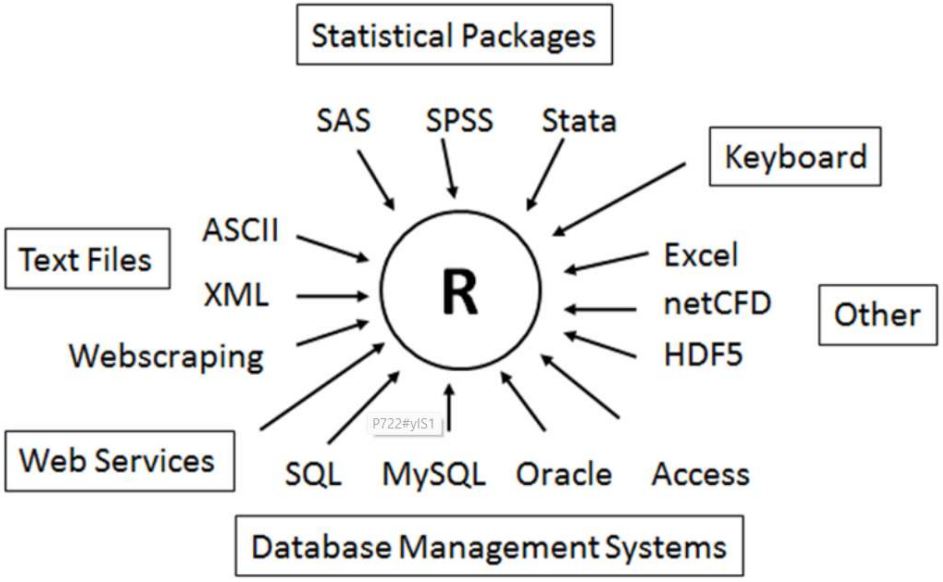
¿Cuáles son las ventajas de R?

- Dispone de un rico ecosistema de bibliotecas o **paquetes,** en su mayoría libres, que extienden sus capacidades. El Repositorio CRAN contiene más de 12.000 paquetes.



¿Cuáles son las ventajas de R?

- Es **compatible** con cualquier tipo de entrada. Las más comunes son los archivos de texto plano, JSON, Excel o CSV, o de bases de datos como SQL Server, Oracle, MySQL y Postgres. También es posible conectar **R** con software de inteligencia de negocios, herramientas de data mining, nubes (AWS, Google Cloud, Microsoft Cloud), Spark Clusters o Hadoop. Además, existen paquetes que permiten realizar web scraping o trabajar directamente con APIs web y extraer datos de esa manera.



¿Cuáles son las ventajas de R?

- En el campo del **Data Science** y el **Machine Learning**, el manejo de **R** es considerado una herramientas fundamental en la formación multidisciplinaria del científico de datos.
- **Integración** con el lenguaje markdown (Rmarkdown) que permiten combinar texto, código y resultados de la evaluación del código en un único documento.
- **Integración** con LaTeX (Paquete Sweave), que permite integrar código **R** en un documento LaTeX con el propósito de crear documentos dinámicos.
- **Integración** con otros lenguajes de programación como C, C++ o Fortran para tareas de análisis de datos computacionalmente intensivas (alto consumo de recursos como CPU y RAM)
- Es muy apreciado en la investigación científica de alto nivel por sus facilidades de **reproducibilidad**. La reproducibilidad es la capacidad de un ensayo o experimento de ser reproducido o replicado por otros, en particular, por la comunidad científica. Este es un requisito cada vez más frecuente para incrementar la calidad e impacto de las publicaciones científicas y organismos patrocinadores de la investigación.

Otras características importantes

- **R** tiene una amplia colección de conjuntos de datos (**datasets**) incorporados (**built-in**) que se distribuyeron originalmente desde las primeras versiones y posteriormente en algunos de los paquetes complementarios. El objetivo es hacer que estos datos sean más accesibles para la enseñanza y el desarrollo de software estadístico. Algunos de los dataset más usados son:
  - **Mtcars**: Este conjunto de datos contiene pruebas en carretera para 32 automóviles (modelos de 1973-74), entre la que podemos encontrar el consumo de combustible, así como aspectos de su diseño y rendimiento. Los datos fueron extraídos de la revista Motor Trend US de 1974.
  - **Iris**: contiene las medidas en centímetros de las variables: longitud de los sépalos, anchura de los sépalos, longitud de los pétalos y anchura de los pétalos de 50 flores de cada una de las 3 especies de iris. El dataset está basado en los datos usados por Ronald Fisher en su artículo de 1936.
  - **PlantGrowth**: Contiene los resultados obtenidos de un experimento para comparar los rendimientos (medidos por el peso seco de las plantas) obtenidos bajo una condición de control y dos tratamientos diferentes.
  - **USArrests**: Este conjunto de datos contiene estadísticas, en arrestos por cada 100.000 residentes por asalto, asesinato y violación en cada uno de los 50 estados de EEUU en 1973.
- En la página web https://vincentarelbundock.github.io/Rdatasets/datasets.html puede verse todos los dataset built-in en **R**.

Otras características importantes

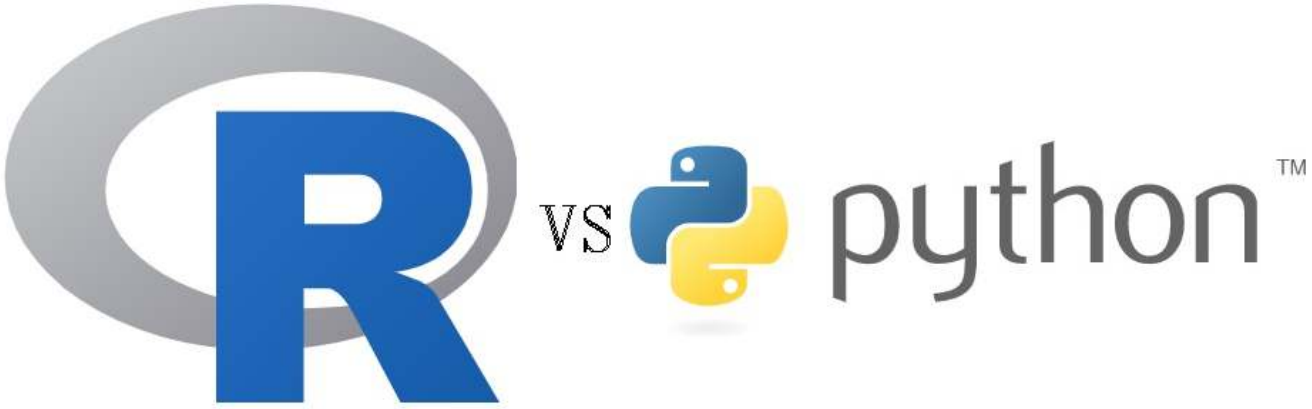
- **Computación distribuida**: La computación distribuida es un modelo en el que los componentes de un sistema de software se comparten entre varios ordenadores para mejorar la eficiencia y el rendimiento. En noviembre de 2015 se publicaron dos nuevos paquetes **ddR** y **multidplyr** utilizados para la programación distribuida en **R**.
- **Computación paralela y de alto rendimiento (HPC)**: Los requerimientos computacionales para el análisis de datos experimentales masivos o Big Data (por ejemplo, en Genómica) y la necesidad urgente de reducción sustancial de los tiempos computacionales, han orientado el desarrollo de **R** hacia paquetes que permiten la ejecución paralela en ambientes multiprocesador. En CRAN Project existen paquetes como **foreach** y **doParallel**, entre otros.

Derivados de R

- **Bioconductor** - https://www.bioconductor.org/ - Herramienta basada en **R** para el análisis riguroso de ensayos biológicos, en particular análisis genómicos.
- **Oracle Machine Learning for R** - https://www.oracle.com/database/technologies/datawarehouse-bigdata/oml4r.html - Es un componente de Oracle Advanced Analytics Option de Oracle Database Enterprise Edition.
- **Microsoft R Application Network** - https://mran.microsoft.com/ - Se puede integrar con Azure.
- **Rapporter** - https://github.com/Rapporter/rapport - Reportes estadísticos desde la nube.
- **Google Collab**: Permite ejecutar R en la nube de Google.

¿R o Python?

R es más amplio que **Python** en la mayoría de las áreas de investigación matemática y estadística. **Python** es más amplio que **R** en Deep Learning, Data Science y en poner modelos en producción, dado que **Python** es un lenguaje multipropósito, no como **R** que nació específicamente para el análisis estadístico.



UNIVERSIDAD YACHAY TECH

SCHOOL OF MATHEMATICAL AND COMPUTATIONAL SCIENCES

FIN

Gustavo A. Colmenares gcolmenares@yachaytech.edu.ec / gcolmena@gmail.com