Manual para el curso-taller

Cómo manipular y limpias bases de datos usando R

Elaborado por Geovana Carreño-Rocabado, PhD.

1 Qué hay en este manual y para que sirve?

Este es un manual en construcción y se irá mejorando con las contribuciones de la comunidad de usuarios R. El objetivo del manual es el de facilitar el uso del programa R para la manipulación y manejo de bases de datos. En la primera parte se presenta un resumen de la importancia y utilidad de este software. Dado que se usará la interfase de Rstudio, se presenta el software y su instalación.

2 Que es el lenguaje R para programación?

R es un entorno y un lenguaje para manipulación de bases de datos, análisis estadísticos y gráficos que fue creado por Ross Ihaka y Robert Gentle (Ihaka and Gentleman 1996). R esta disponible como un software libre (sin costo) bajo los términos de Free Software Foundation's GNU General Public License en la fuente del código. Es un lenguaje orientado a objetos, lo que significa que las variables, datos, funciones, resultados, etc., se guardan en la memoria activa del computador en forma de objetos con un nombre específico. El usuario puede modificar o manipular estos objetos con operadores (aritméticos, lógicos, y comparativos) y funciones (que a su vez son objetos) es decir crea objetos y los manipula con un objetivo. También es un lenguaje de programación de funciones. R es muy flexible y en muchos casos muy comprensible para entender el procesos y el producto.

3 Por qué usar R

Hay muchas razones para iniciar el entrenamiento en el uso de R. Entre ellas que es un software libre, que cuenta con una gran comunidad de apoyo y que esta en permanente mejora. Específicamente para la manipulación y limpieza de bases de datos, trabajar con R tiene las ventajas de:

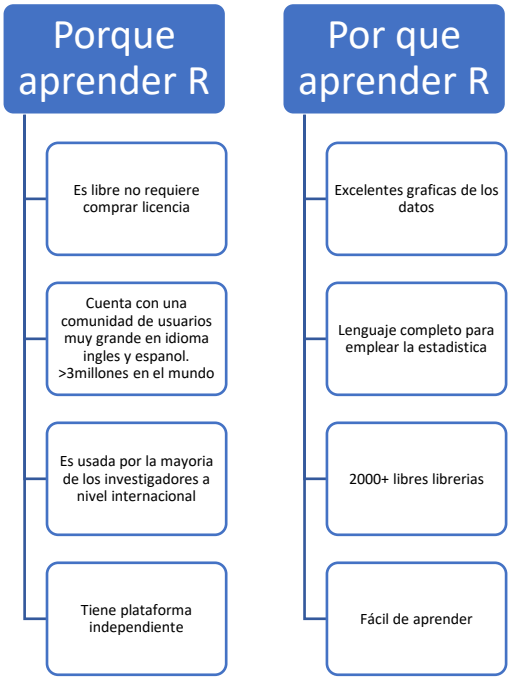
Los cambios no se realizan en la base misma, sino en objetos creados en base a ella

Se lleva un registro de los cambios lo que ayuda a identificar errores y facilita su corrección

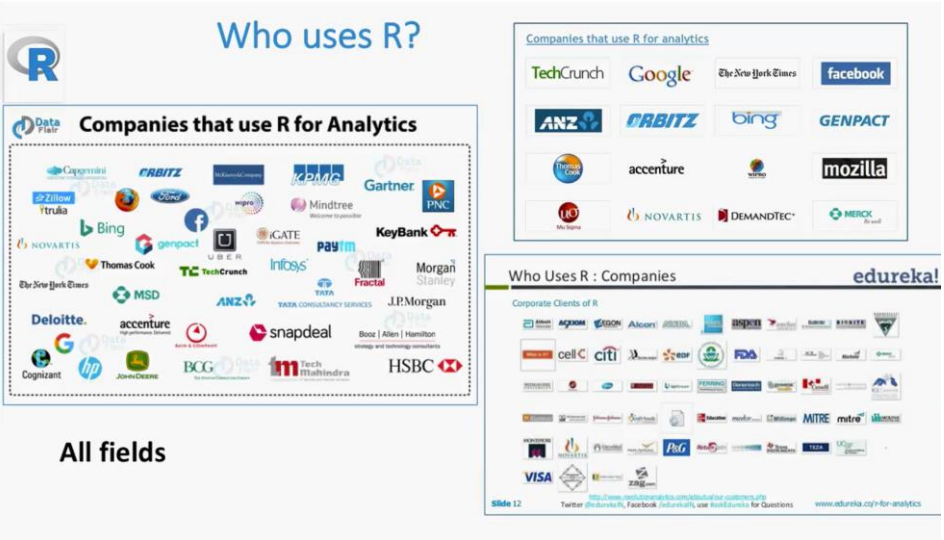
Los cambios se van guardando en diferentes etapas, esto permite hacer pausas en el trabajo y retomar revisando que se hizo antes.

Se puede combinar una serie de bases para verificar relaciones o identificar errores

Se puede usar una serie de graficas que facilitan entender las tendencias de los datos e identificar problemas



Muchas conocidas compañías usan R para analizar sus datos.



Si realmente quiere aprender R se debe apoyar en una comunidad existente

4 Que es R studio

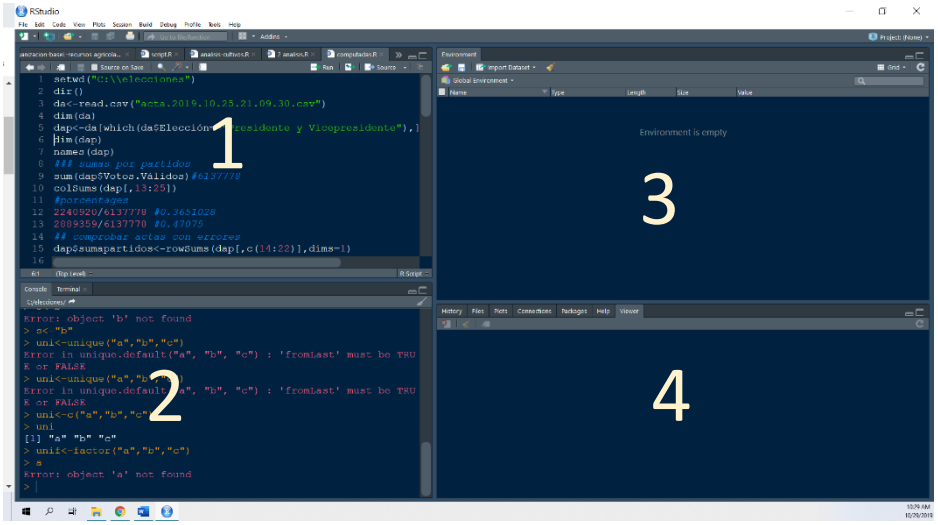
En una interfase para usar R más preferida por los usuarios de R. Es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R. Esta integrado y es simple, en una sola plataforma se puede ver y trabajar varias opciones. En este curso usaremos RStudio como interfase para manejar R.

RStudio tiene cuatro ventanas con las siguientes funciones.

1. Editor de códigos (scripts) y visor de datos

2. Consola de R que muestra las salidas de los scripts

3 y 4 Espacio para ver ambiente de trabajo (objetos que están activos), historia, archivos, graficas, conexiones, paquetes, ayuda, build, VCS y Viewer.



5 Donde empiezo?

Para iniciar a usar R primero debemos instalarlo en la computadora. Después debemos instalar RStudio.

5.1 Bajar e instalar R

1. Abrir un buscador de internet e ir a la dirección www.r-project.org
2. Hacer click al link de “download R” en el medio de la pagina bajo “Getting started”
3. Escoger el Mirror que son lugares donde se encuentran copias de las librerias y demás archivos para R. Se puede escoger Brasil, Chile, EEUU alguno que este mas cerca a nuestro país.

- Depende de tu Sistema operativo escoger uno de los tres links para bajar R "Download R for Linux" o "Download R for (Mac) OS X" o Download R for Windows".

4. Click en " **install R for the first time**"
5. Click en Download R 3.6.1 for Windows, y ver la carpeta donde se bajará este archivo, o dar una dirección al archivo que bajaran (En caso de tener otro sistema operativo, buscar el que corresponda y bajarlo)
6. Abrir el archivo que se bajo y seguir las instrucciones para su instalación. Si tienes dudas en esta instalación pueden revisar el manual de Collaton 2014, en las paginas 6 a 10 tiene los detalles, solo que esto muestra una versión anterior del R. Sin embargo la instalación es similar.

7. Ahora que R ya esta instalado puedes abrirlo para que compruebes que esta bien y esa es la ventana que encontraras (Anexo. Ventana de R). Cierra el programa e instala Rstudio

5.2 Bajar e instalar R Studio

1. Ir a www.rstudio.com y hacer click "Download Rstudio"
2. Click en "Download Rstudio Desktop"
3. Click en la versión recomendad para tu sistema, guardar el archivo en la computadora, hacer doble click para abrirlo, luego
4. Buscar el archivo y abrirlo, seguir los pasos para la instalación

Si tienes dudas en esta instalación pueden revisar el manual de Collaton 2014, en las paginas 6 a 10 tiene los detalles, solo que esto muestra una versión anterior del R. Sin embargo la instalación es similar.

6 Paquetes de R (R packages)

6.1 Que son ellos?

Los paquetes en R son colecciones de funciones y conjunto de datos desarrollados por la comunidad. Estos incrementan la potencialidad de R mejorando las funcionalidades baseen R, o añadiendo de nuevas. Continuamente son desarrollados nuevos paquetes y dependiendo de lo que necesitamos hacer o como lo necesitamos hacer necesitaremos diferentes paquetes. R viene con un conjunto de paquetes.

6.2 Quienes los desarrollan?

Cualquier usuario de la comunidad de R que quiera contribuir a este proyecto y que tenga los conocimientos para ello.

6.3 Como instalarlos?

1. En Rstudio editor de codigos

Install.packages("nombre del paquete")

De la pestana tools-> Install packages-> escribir el nombre de paquete->install

6.4 Como usar los paquetes en R

Cargar el paquete

library(“nombre del paquete”)

El paquete está listo para usar.

7 Cómo conseguir ayuda

1. Usar las hojas de ayuda de diferentes paquetes de “cheat sheets” de la pagnia web de Rstudio

https://rstudio.com/resources/cheatsheets/

2. En la consola escribir ??funcion

8 Tipo de datos

8.1 Vector

a<-c(1,2,3,4,5.3,6,-2) # vector numeral

b<-c(“uno”,”dos”,”tres”) # vector de cartacteres

c<-c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE) # vector logicos

8.2 Matrices

Y<-matrix(1:20, nrow=5,ncol=4) # genera 5x4 matriz numerica

Array. Son similares a las matrices pero pueden tener mas de dos dimensiones

8.3 Listas

a<-list(names="Ana", mynumbers=a,mymatrix=y, age=5.3)## genera una lista de 3 elementos

8.4 Factores

genero<-c(rep("hombre",20), rep("mujer",30))

genero<-factor(genero)

levels(genero)

9 Apoyo para trabajando con R

1. Cheat sheet de cada paquete
2. ?? nombre de la función en consola de R
3. Libros de R
4. Data Camp cursos
5. Comunidad de usuarios de R
6. Buscar en Google la duda, mejor en ingles

10 Buenas practicas para programa en R

1. Siempre crear un nuevo folder para cada proyecto si es aplicable
2. Siempre crear “nuevo proyecto” por proyectos
   File->new directory -> new Project -> nombre del directorio y lugar para guardar->create new Project
3. Indicar el directorio en que se trabajara el ejercicio
4. Usar cortas rutas de guardar
5. Nombrar a los objetos con nombre cortos pero comprensivos. Tener un sistema para poner los nombres.
6. Usar comentarios al escribir los códigos con el signo “#”.
7. Escribe tus códigos los mas sistemática y ordenadamente posible
8. Ser fiel a R y practicar, practicar y practicar y practicar.
9. Ten un cuaderno de notas de tus códigos mas relevantes a cada análisis

11 R es sensible a

1. Mayúsculas
2. Acentos
3. Espacios
4. Celdas vacías

12 Bibligrafía

Ihaka, R., and R. Gentleman. 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 5(3):299–314.