

BIG DATA & DATA SCIENCE

VÍCTOR LEIVA

www.victorleiva.cl

ESCUELA DE INGENIERÍA INDUSTRIAL

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

AGENDA

BUSINESS INTELLIGENCE (BI)

BIG DATA

DATA MINING

DATA SCIENCE

EXTRACT, TRANSFORM, LOAD (ETL)

PRESENTACIÓN DE CASOS

SOFTWARE R

ANÁLISIS DE CASOS EN R

USE

CR

QUÉ ES R

R ES UN SOFTWARE ESTADÍSTICO (www.r-project.org) IDEADO POR IHAKA & GENTLEMAN (1996)*.

R ES LA VERSIÓN GRATUITA DEL SOFTWARE ESTADÍSTICO S ([s-plus](http://s-plus.com)) Y ESTÁ EN LA LÍNEA DE LOS FREEWARE GNU (www.gnu.org).

EL CÓDIGO FUENTE DE R ESTÁ ESCRITO EN LENGUAJE C Y ALGUNAS RUTINAS EN LENGUAJE FORTRAN. ENTONCES, R PUEDE SER RECONOCIDO COMO UN LENGUAJE DE PROGRAMACIÓN COMPUTACIONAL.

R ES UN LENGUAJE “INTERPRETADO” (COMO JAVA) Y NO “COMPILADO” (COMO C O FORTRAN). ESTO SIGNIFICA QUE LOS COMANDOS ESCRITOS EN EL TECLADO SE EJECUTAN DIRECTAMENTE SIN NECESIDAD DE UN COMPILADOR.

*IHAKA, R., GENTLEMAN, R. (1996) R: A LANGUAGE FOR DATA ANALYSIS AND GRAPHICS. JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS 5:299–314.

QUÉ ES R

R ES UN LENGUAJE MATRICIAL ORIENTADO A OBJETOS: ESTO SIGNIFICA QUE LAS VARIABLES Y FUNCIONES, Y LOS DATOS Y RESULTADOS, ETCÉTERA, SE GUARDAN EN LA MEMORIA ACTIVA DEL COMPUTADOR (RAM) EN FORMA DE OBJETOS CON UN NOMBRE ESPECÍFICO.

EL USUARIO PUEDE MODIFICAR O MANIPULAR LOS OBJETOS R CON OPERADORES (ARITMÉTICOS, LÓGICOS O COMPARATIVOS) Y FUNCIONES (QUE A SU VEZ SON TAMBIÉN OBJETOS).

R ES UN PAQUETE DE SOFTWARE NO COMERCIAL DE CÓDIGO ABIERTO PARA COMPUTACIÓN ESTADÍSTICA Y GRÁFICOS QUE SE PUEDE DESCARGAR GRATUITAMENTE DESDE <http://CRAN.R-project.org>.

R TIENE UN MANEJO DE DATOS MUY SIMPLE. R ES MUY VERSÁTIL PARA CONSTRUIR GRÁFICOS.

R CONSTA DE PAQUETES “BASE” Y DE PAQUETES ADICIONALES QUE EXTIENDEN SU FUNCIONALIDAD.

QUÉ ES R

INSTALAR PAQUETES ADICIONALES EN **R** ES SIMPLE. SE PUEDE HACER USANDO LAS FUNCIONES **R: `install.packages`** Y **`update.packages`** O USANDO EL GUI DE SUS VENTANAS.

R PERMITE IMPLEMENTAR NUEVOS MÉTODOS ESTADÍSTICOS QUE NO ESTÁN DISPONIBLES EN OTROS SOFTWARE A TRAVÉS DE PAQUETES.

R ES UN LENGUAJE PROGRAMACIÓN. SIN EMBARGO, HAY PAQUETES QUE NOS PERMITEN USAR UN INTERFAZ GRÁFICA, A MODO DE VENTANAS (COMO EN EL CASO DE **STATA** O **SPSS**, POR EJEMPLO), PARA REALIZAR ALGUNOS ANÁLISIS DE UNA MANERA MÁS ACCESIBLE Y SIN CONOCER NADA DE **R**.

EL PAQUETE **R** MÁS FAMOSO PARA ESTA INTERFAZ GRÁFICA ES **Rcmdr** (**Rcommander**).

VENTAJAS Y DESVENTAJAS DE R

VENTAJAS

SOFTWARE DE DISTRIBUCIÓN **GRATUITA**, **LIGERO** Y **FÁCIL DE INSTALAR**.

CUALQUIER USUARIO DE **R** PUEDE CREAR SUS PROPIAS FUNCIONES Y ANÁLISIS PERSONALIZADOS.

GRAN CAPACIDAD PARA ALMACENAR Y MANIPULAR BASES DE DATOS.

ACCESO A UNA GRAN FUENTE DE INFORMACIÓN EN INTERNET (BLOG, FOROS, MANUALES, ENTRE OTROS)

SOBRE PROGRAMACIÓN EN **R**.

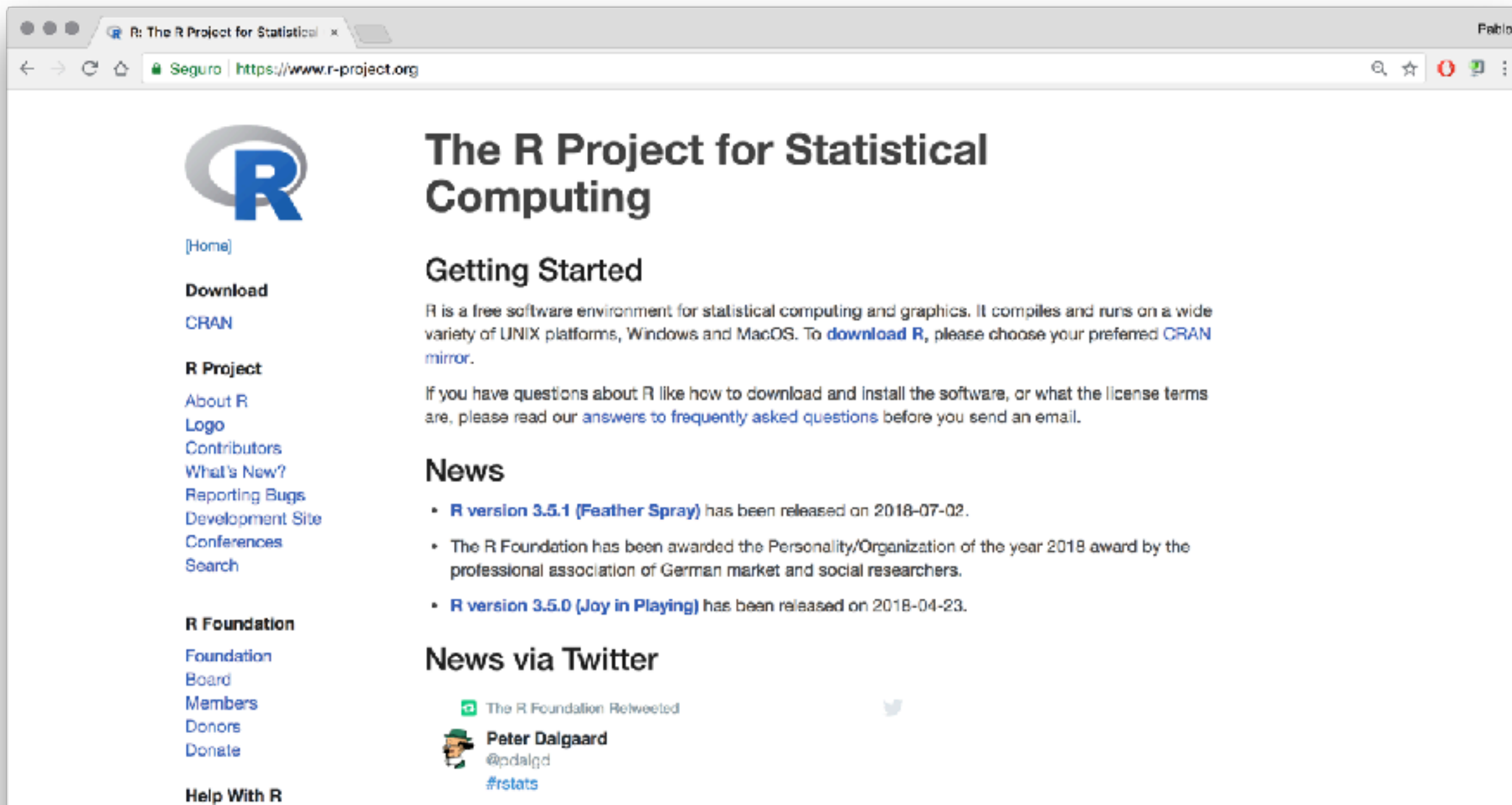
DISPONIBLE EN LINUX, MAC-OS Y WINDOWS.

DESVENTAJA

NO ES FÁCIL DE USAR PARA QUIENES NO TIENEN CONOCIMIENTOS BÁSICOS EN PROGRAMACIÓN, PERO HAY ALTERNATIVAS INTERACTIVAS.

DESCARGANDO R

DESCARGAR EL SOFTWARE R DESDE <http://www.r-project.org>.
O DESDE <http://CRAN.R-project.org>.



The screenshot shows a web browser window with the address bar displaying <https://www.r-project.org>. The page features the R logo on the left, a navigation menu with links like [Home], Download, CRAN, R Project, About R, Logo, Contributors, What's New?, Reporting Bugs, Development Site, Conferences, Search, R Foundation, Foundation, Board, Members, Donors, Donate, and Help With R. The main content area is titled 'The R Project for Statistical Computing' and includes a 'Getting Started' section with text about R being a free software environment for statistical computing and graphics, and a 'News' section with three bullet points: 'R version 3.5.1 (Feather Spray)' released on 2018-07-02, 'The R Foundation' awarded the Personality/Organization of the year 2018 award by the professional association of German market and social researchers, and 'R version 3.5.0 (Joy in Playing)' released on 2018-04-23. At the bottom, there is a 'News via Twitter' section showing a tweet from Peter Dalgaard (@pdalgd) with the hashtag #rstats.

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- **R version 3.5.1 (Feather Spray)** has been released on 2018-07-02.
- The R Foundation has been awarded the Personality/Organization of the year 2018 award by the professional association of German market and social researchers.
- **R version 3.5.0 (Joy in Playing)** has been released on 2018-04-23.

News via Twitter

The R Foundation Retweeted

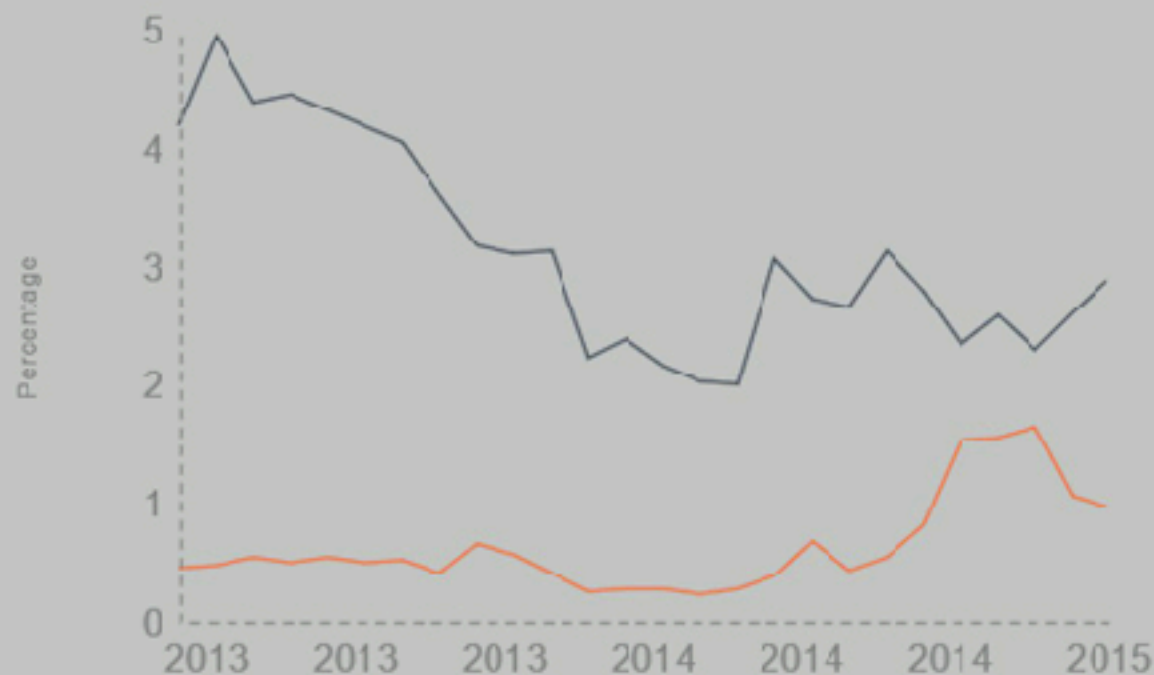
Peter Dalgaard
@pdalgd
#rstats

R VS PYTHON

R and Python: The Numbers

Popularity Rankings

R and Python's popularity between 2013 and February 2015 (Tiobe Index)



Jobs And Salary?

2014 Dice Tech Salary Survey:
Average Salary For High Paying Skills and Experience



\$ 115,531

Python

R

Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)



\$94,139

It's a tie!
It's up to you, the data scientist,
to pick the language that best fits your needs.
The following questions can guide you in your decision.

1

What problems do you want to solve?

2

What are the net costs for learning a language?*

* it will cost time to learn a new system that is better aligned for the problem you want to solve, but staying with the system you know may not be made for that kind of problem.

3

What are the commonly used tool(s) in your field?

4

What are the other available tools in your field and how do these relate to the commonly used tool(s)?

SOFTWARE EN DATA SCIENCE

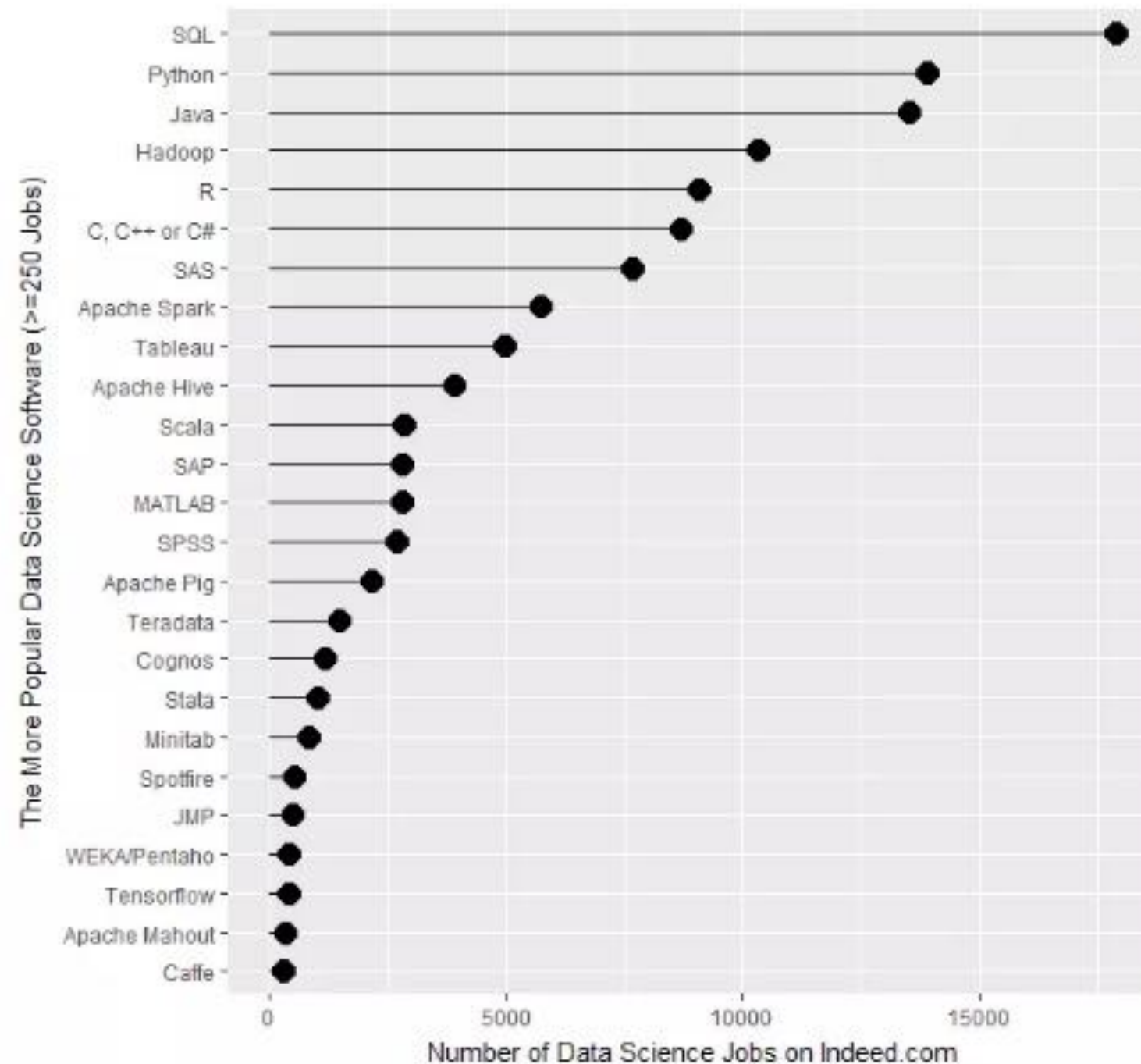
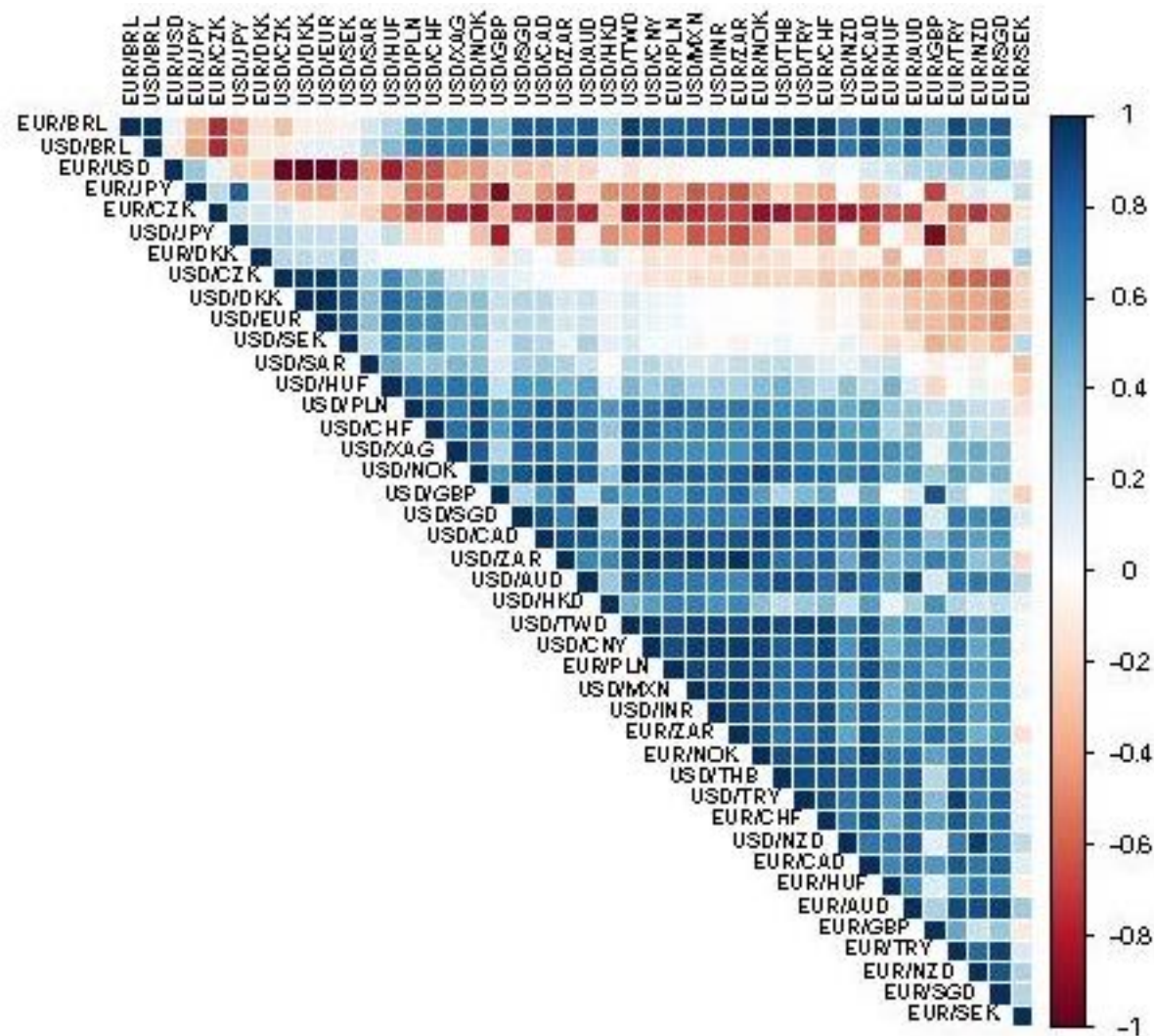


Figure 1a. The number of data science jobs for the more popular software (those with 250 jobs or more, 2/2017).

TASAS DE CAMBIO EN R



SE QUIERE CONOCER LA CORRELACIÓN ENTRE LAS TASAS DE CAMBIO USD/BRL Y EUR/BRL CON RESPECTO A LAS PRINCIPALES DIVISAS Y TAMBIÉN ENTRE ELLAS. CON **R** SE PUEDE ACCEDER A BASES DE DATOS EN LÍNEA DESDE PÁGINAS WEB. CONSIDERE LA SERIE ANUAL (DIARIA) ENTRE EL **4 DE ABRIL DE 2015** Y EL **4 DE ABRIL DE 2016** DE DIFERENTES DIVISAS Y TASAS DE CAMBIO TOMADA DESDE [WWW. OANDA.COM](http://www.oanda.com). LOS VALORES EN ROJO CORRESPONDEN A UNA **CORRELACIÓN NEGATIVA**, MIENTRAS QUE LOS VALORES EN AZUL A UNA **CORRELACIÓN POSITIVA**.

SI NO SE USA **R**, HABRÍA QUE HACER EL **QUERY** DE A PARES DE DIVISAS DEL ÚLTIMO AÑO, UNA A UNA, EN LA WEB: **¿CUÁNTO TIEMPO HABRÍAMOS DEMORADO?**

OANDA

Abra una cuenta Pruebe una demostración gratuita Inicie sesión Introduzca pal

Por qué OANDA Plataforma Mercado Aprender Noticias y análisis Ayuda Convertidor de divisas Soluciones para empresas

Convertidor de divisas Herramientas de Cambio Aplicaciones Móviles

Tipos de cambio históricos

Convertidor de divisas Tipos de cambio históricos Tipos de cambio en directo

Tengo esta divisa: US Dólar USD Quiero esta divisa: EUR

PERIODO: 30 días 10 mar 2016 00 abr 2016 TIPO INTERBANCARIO: +/- 0%

PRECIO: Compra VALORES: tipos de cambio FRECUENCIA: Diario

TRY API... Exchange Rate Feeds Mobile Currency Apps Add Tools to Your Site

CON **R** PODEMOS RESPONDER PREGUNTAS DEL TIPO:

¿CÓMO EXTRAEMOS LAS BASES DE DATOS DESDE UN SERVIDOR WEB?

¿CÓMO TRANSFORMAMOS LAS BASES DE DATOS EXTRAÍDAS EN UNA SOLA BASE DE DATOS?

¿CÓMO CARGAMOS LA BASE DE DATOS TRANSFORMADA PARA HACER ANALYTICS?

EL CASO DE OANDA ES ALGO BÁSICO EN **R** (NO USA PAQUETES). SE PUEDE ACCEDER A OANDA ONLINE Y ACTUALIZAR CADA DÍA. SIN **R** SE TENDRÍAMOS QUE CALCULAR UNA MATRIZ DE 42×42 .

ETL EN R

R ES CAPAZ DE **EXTRAER, TRANSFORMAR Y CARGAR** LOS DATOS A SER ANALIZADOS. COMO **R** ES UN SOFTWARE GRATUITO, SE CREARON PAQUETES PARA CADA PARTE DE ETL CON APLICACIONES EN BUSINESS INTELLIGENCE. ALGUNOS DE ESTOS PAQUETES SON:

EXTRAER Y CARGAR: `RODBC`, `DBI`, `RJDBC`.

TRANSFORMAR: `data.table`, `dplyr`.

ESTO REFLEJA LAS DIVERSAS OPCIONES QUE **R** OFRECE PARA RESOLVER EL PROBLEMA DE ETL. NO EXISTE UN MÓDULO ÚNICO PARA CADA ETAPA DE ETL. ASÍ, **R** ES CAPAZ DE ASUMIR EL ROL DE MOTOR DE PROCESOS RELACIONADOS A **BUSINESS INTELLIGENCE**.

ETL EN R, PASO A PASO

1. EXTRAER HACIENDO UN QUERY AL WEBSITE DONDE ESTÁN LOS DATOS MEDIANTE UN CICLO FOR DE **R**. CADA VEZ QUE SE EXTRAER UNA BASE DE DATOS DESDE EL SERVIDOR, SE GUARDA EN UN ARCHIVO EN UNA CARPETA DE TRABAJO.
2. CARGAR CADA BASE DE DATOS EXTRAÍDA EN FORMATO DELIMITADO POR COMAS (*.CSV), TAMBIÉN MEDIANTE UN CICLO FOR.
3. TRANSFORMAR LAS BASES DE DATOS CARGADAS EN UN SOLO CONJUNTO DE DATOS PARA HACER ANALYTICS.
4. REALIZAR LAS ANALYTICS PERTINENTES.
5. EN EL CASO DE OANDA, LA ANALYTIC ES EL GRÁFICO DE UNA MATRIZ DE CORRELACIÓN, QUE SE REALIZA CON EL COMANDO `corrplot` DE **R**.

R ES UNA POTENTE HERRAMIENTA DE ESTADÍSTICA COMPUTACIONAL PARA **BUSINESS INTELLIGENCE**.

COMPLEMENTOS DE R

EXISTEN ALGUNOS PROGRAMAS COMPUTACIONALES QUE COMPLEMENTAN EL R. **Tinn-R** ES UN EDITOR GRATUITO PARA ESCRIBIR CÓDIGOS Y RUTINAS EN R DE FORMA AMIGABLE.

R-enterprise (REVOLUTIONS) ES OTRO COMPLEMENTO DE R QUE ESTÁ LLAMANDO LA ATENCIÓN DE LAS EMPRESAS. ÉSTE ES UN SOFTWARE PAGADO QUE USA R PARA RESOLVER PROBLEMAS DE BIG DATA Y BUSINESS INTELLIGENCE. MICROSOFT COMPRÓ **R-enterprise** EN 2015.

Rcommander (MEDIANTE EL PAQUETE **Rcmdr**) ES UNA INTERFAZ GRÁFICA GRATUITA DE R, FÁCIL E INTUITIVA DE USAR. R ES EL MOTOR DE ANÁLISIS Y **Rcmdr** LA MANERA INTERACTIVA PARA REALIZARLO. **Rcommander** NO NECESITA TENER CONOCIMIENTOS DE PROGRAMACIÓN Y PERMITE CARGAR DATOS DESDE OTROS SOFTWARE.

R-Studio ES UN COMPLEMENTO GRATUITO DE R QUE UTILIZA LA MEMORIA GRÁFICA DEL COMPUTADOR PARA FACILITAR LAS INTERACCIONES CON R. **R-Studio** Y **Rcmdr** SON SIMILARES EN SU ENTORNO GRÁFICO E IMITAN LOS SOFTWARE PAGADOS CLÁSICOS (COMO **SPSS** O **STATA**) CON MENÚS Y MÓDULOS INTERACTIVOS.

INTERACTUANDO CON R

LA AYUDA EN LÍNEA DE **R** PRESENTA INFORMACIÓN ÚTIL SOBRE CÓMO UTILIZAR SUS FUNCIONES. POR EJEMPLO, PARA SABER QUÉ HACE LA FUNCIÓN `lm`, ESCRIBIR: `?lm`, O BIEN `help(lm)`

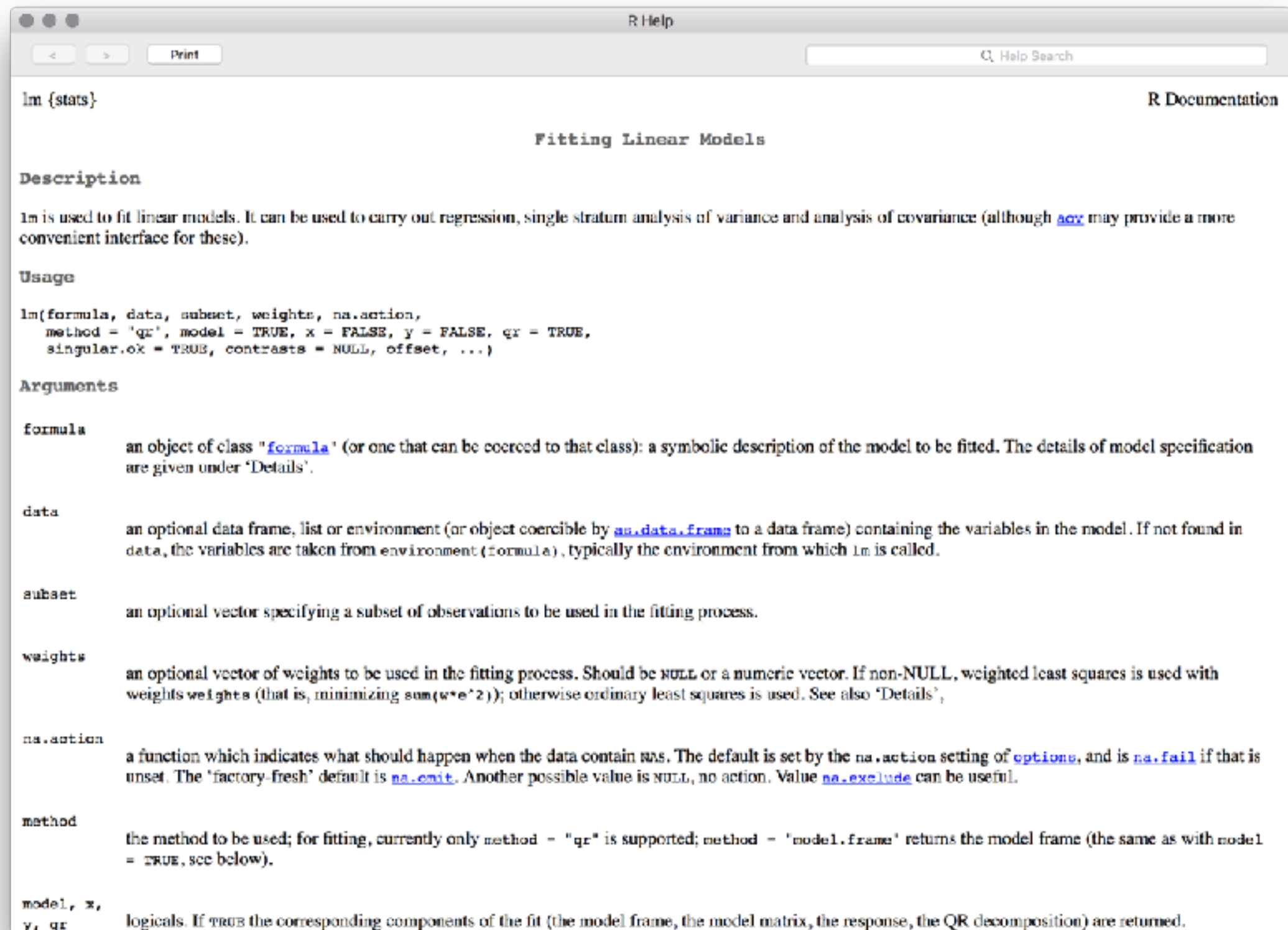
AL LLAMAR LA AYUDA, SE ABRE UNA PÁGINA WEB CON TODA LA INFORMACIÓN SOBRE LA FUNCIÓN. LA PÁGINA PRINCIPAL DE LA AYUDA SE LLAMA ESCRIBIENDO: `help.start()`

CUANDO NO CONOCEMOS CON EXACTITUD EL NOMBRE DE UNA FUNCIÓN, SE PUEDE UTILIZAR `apropos()`. ÉSTA ENCUENTRA TODAS LAS FUNCIONES CUYO NOMBRE CONTIENE UNA PALABRA DADA EN EL PAQUETE BASE O EN LOS PAQUETES CARGADOS EN MEMORIA. POR EJEMPLO:

```
apropos("prod")  
[1] "crossprod" "cumprod"   "prod"      "tcrossprod"
```

MAYOR DETALLE EN: <https://renlinea.wordpress.com>

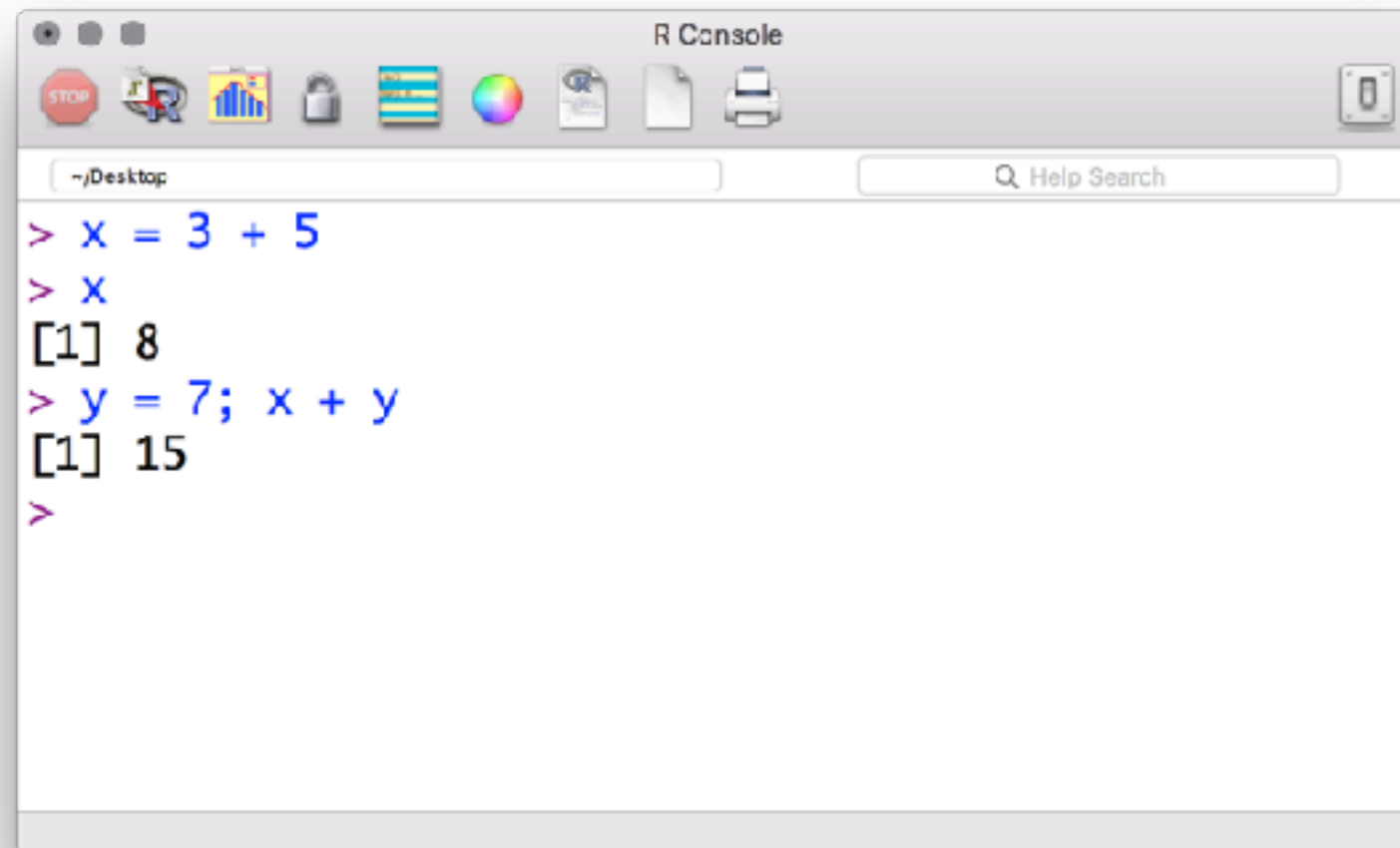
INTERACTUANDO CON R



INTERACTUANDO CON R

R FUNCIONA DE FORMA INTERACTIVA, UTILIZANDO UN MODELO DE PREGUNTAS Y RESPUESTAS:

1. PARTIR CON R.
2. TIPEAR UN COMANDO EN LA CONSOLA DE R Y PRESIONAR «ENTER».
3. R ESPERA ENTONCES POR UN INPUT MÁS.
4. TIPEAR `q()` PARA SALIR DE R.



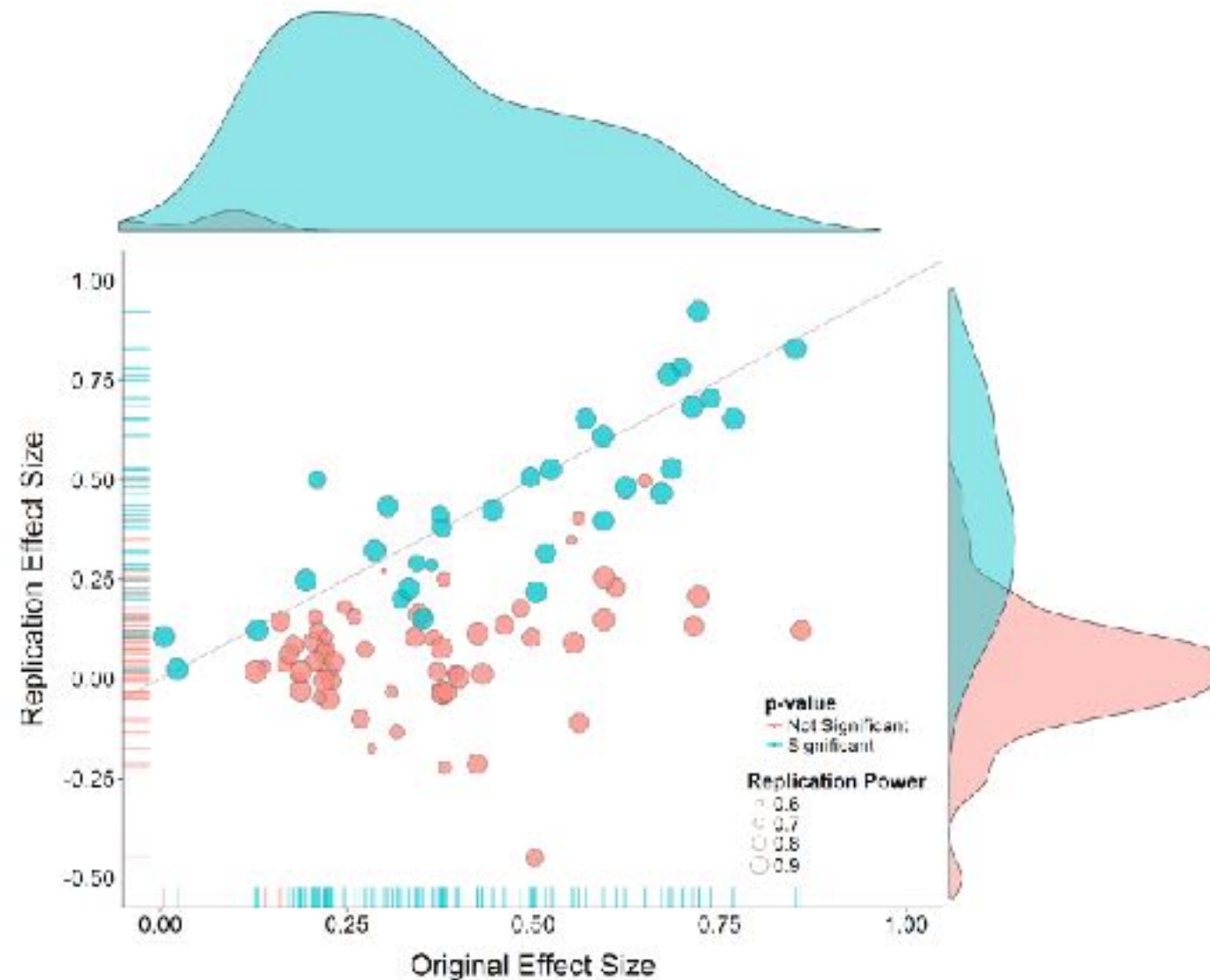
```
R Console
~/Desktop
[1] 8
[1] 15
>
```

The screenshot shows the R Console window with a toolbar at the top containing icons for stopping, saving, plotting, locking, editing, color, help, and printing. Below the toolbar is a search bar labeled 'Help Search'. The main area of the console displays the following text:

```
> x = 3 + 5
> x
[1] 8
> y = 7; x + y
[1] 15
>
```

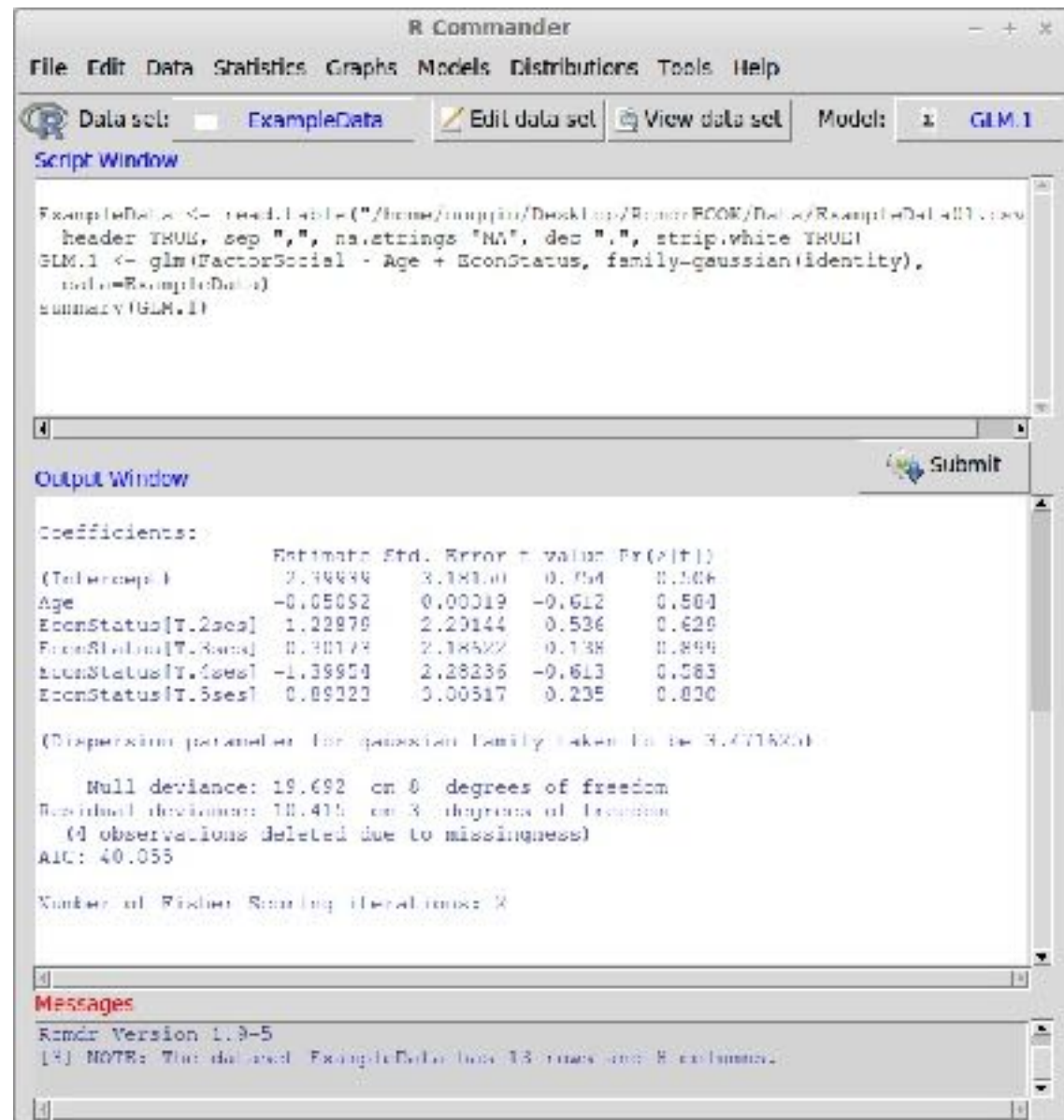
GRÁFICOS EN R

CON PACIENCIA, CON  SE PUEDE LOGRAR LO SIGUIENTE:



VER OTROS GRÁFICOS Y EL CÓDIGO DE ESTE GRÁFICO EN: [HTTPS://GOO.GL/5HZ302](https://goo.gl/5HZ302)

INTRODUCCIÓN A R COMMANDER



R Commander ES UNA INTERFAZ GRÁFICA DE USUARIO.

R Commander ES FÁCIL E INTUITIVO DE USAR.

R Commander ES UNA FORMA INTERACTIVA DE

ANALIZAR DATOS CUYO MOTOR ES R. R Commander

NO NECESITA TENER CONOCIMIENTOS DE PROGRAMACIÓN.

R Commander PERMITE ANALIZAR DATOS PROVENIENTES

DE OTROS SOFTWARE. PARA EJECUTAR R Commander,

PRIMERO INSTALE EL PAQUETE **Rcmdr** Y LUEGO EJECUTE:

library(Rcmdr) EN LA CONSOLA DE R.

<http://www.rcommander.com>

UN POCO DE BIBLIOGRAFÍA

BAESEN, B. (2014)

ANALYTICS IN A BIG DATA WORLD: THE ESSENTIAL GUIDE TO DATA SCIENCE AND ITS APPLICATIONS. WILEY, NEW YORK.

DEAN, J. (2014)

BIG DATA, DATA MINING, AND MACHINE LEARNING: VALUE CREATION FOR BUSINESS LEADERS AND PRACTITIONERS. WILEY, NEW YORK.

DIETRICH, D. (2015)

DATA SCIENCE AND BIG DATA ANALYTICS: DISCOVERING, ANALYZING, VISUALIZING AND PRESENTING DATA. WILEY, NEW YORK.

HASTIE, T., TIBSHIRANI, R. (2016)

THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION. SPRINGER, NEW YORK.

HURWITZ, J., KAUFMAN, M., BOWLES, A. (2015)

COGNITIVE COMPUTING AND BIG DATA ANALYTICS. WILEY, NEW YORK.

PROVOST, F., FAWCETT, T. (2013)

DATA SCIENCE FOR BUSINESS: WHAT YOU NEED TO KNOW ABOUT DATA MINING AND DATA-ANALYTIC THINKING. O'REILLY MEDIA.

WALKOWIAK, S. (2016)

BIG DATA ANALYTICS WITH R. PACKT PUBLISHING.

HURWITZ, J., KAUFMAN, M., BOWLES, A. (2015)

COGNITIVE COMPUTING AND BIG DATA ANALYTICS. WILEY, NEW YORK.

BIG DATA & DATA SCIENCE

VÍCTOR LEIVA

www.victorleiva.cl

ESCUELA DE INGENIERÍA INDUSTRIAL

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO