Data Mining Consulting

R for Bussiness Analytics

Introducción al Análisis Predictivo con R

Mg Jesús Salinas Flores

jesus.salinas@dataminingperu.com

Data Mining Consulting

Presentación del curso

Mg Jesús Salinas Flores

jesus.salinas@dataminingperu.com

Expositor

Ingeniero Estadístico egresado de la Universidad Nacional Agraria La Molina

Mg. en Ingeniería Industrial con especialidad en Gestión Industrial egresado de la Universidad Nacional Mayor de San Marcos

Profesor principal del dpto. de Estadística e Informática en la UNA La Molina

Docente en la maestría de Estadística Aplicada (UNA La Molina) y en la maestría de Ciencias de los Datos (U Ricardo Palma)

Expositor

Áreas de interés

- Análisis multivariado
- Reconocimiento estadístico de patrones

- Jesús Salinas (a)
- Estadística para todos

- Estadística para todos

Dirigido a:

Dirigido a todos aquellos profesionales interesados en ampliar sus conocimientos en herramientas de análisis de datos

Profesionales que se desempeñen como investigadores, analistas de datos

Personal de Procesamiento de Datos, Business Intelligence, Analistas de Datos, etc

Estudiantes de maestría y de pregrado

Programa del curso

I. Introducción al R

- ¿Qué es R?
- Instalación de paquetes
- Comandos Básicos
- Recursos en Línea
- Como funcionan el Manejo de Datos en R
- Manejo de archivos de datos
- Importar. Exportar

II. Gráficos con R

- Manejo y distribución de Gráficos
- Funciones Gráficas
- Parámetros. Comandos de Gráficos
- Uso del paquete ggplot2

III. Business Analytics con R

- Análisis Descriptivo
- Medidas estadísticas
- Medidas de variabilidad
- Medidas de Asimetría

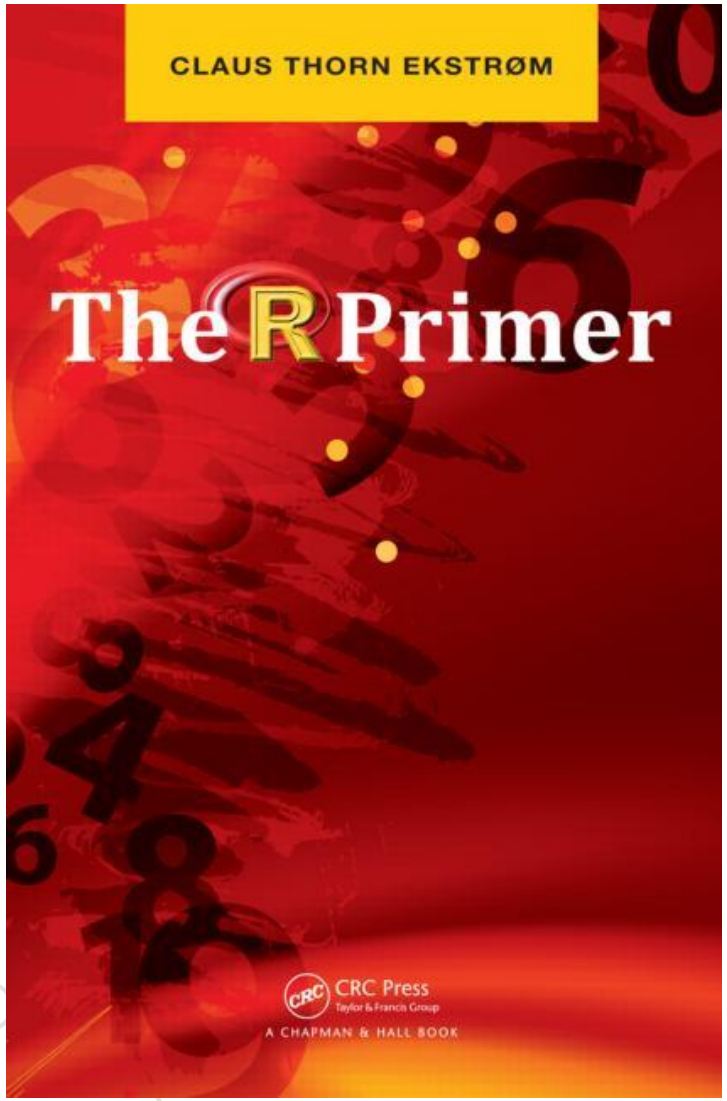
IV. Modelos Predictivos

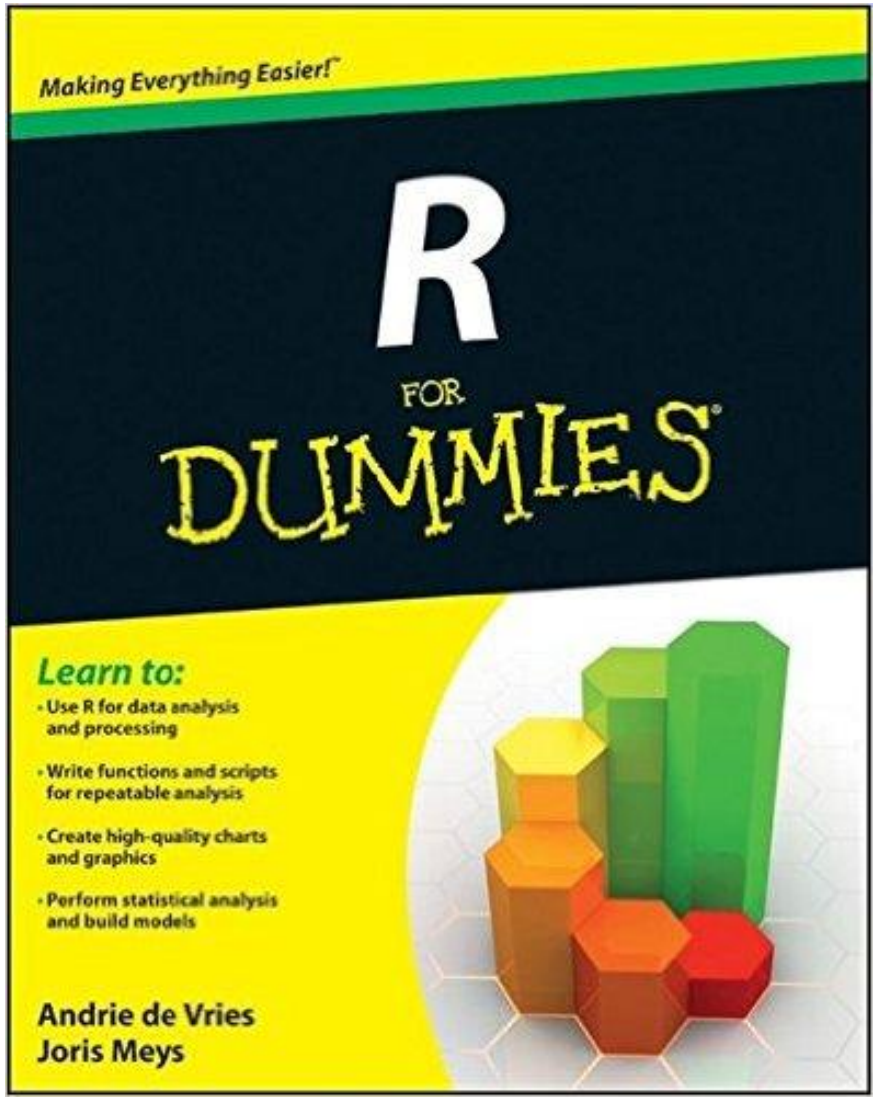
- Análisis Cluster. K-means y algoritmos jerárquicos
- Regresión Lineal. Regresión Múltiple
- Árboles de Clasificación
- Uso de Rattle

Sistema de evaluación

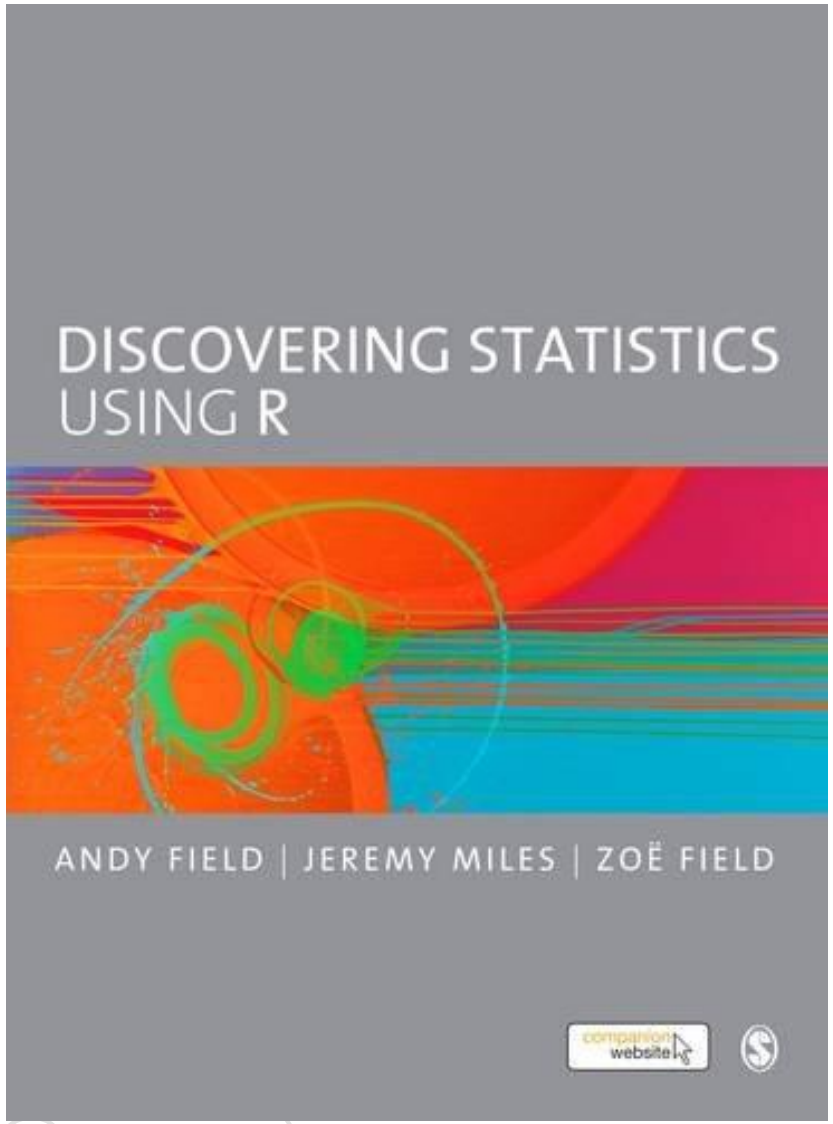
Desarrollo de trabajos en clase
• 100%

Bibliografía del curso

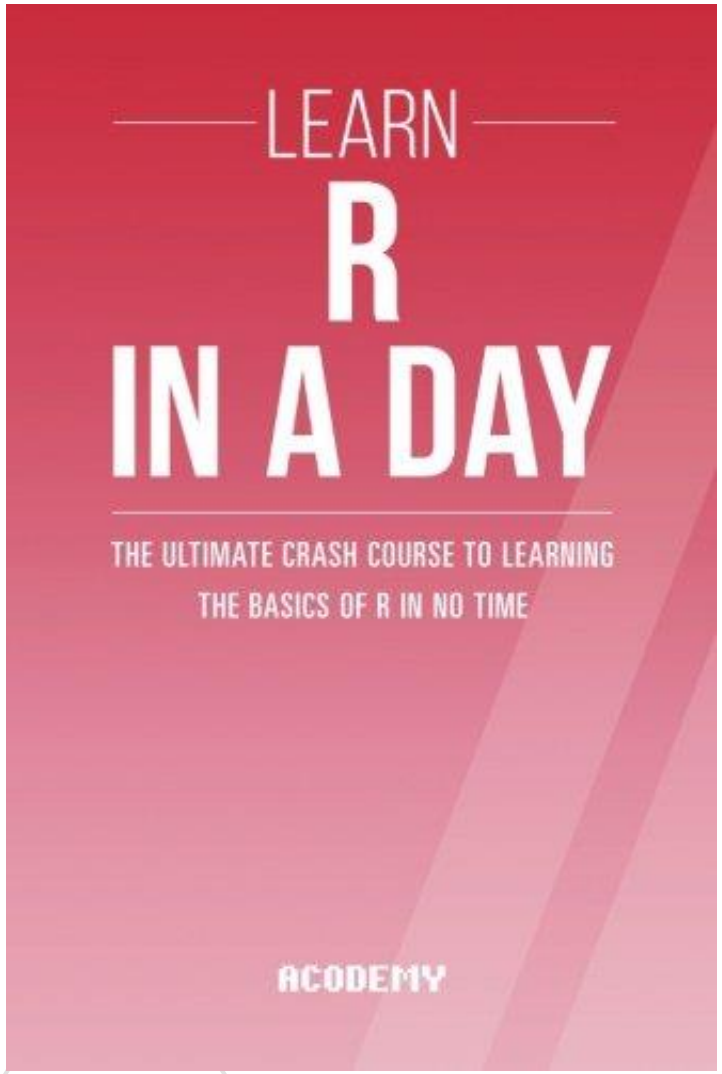


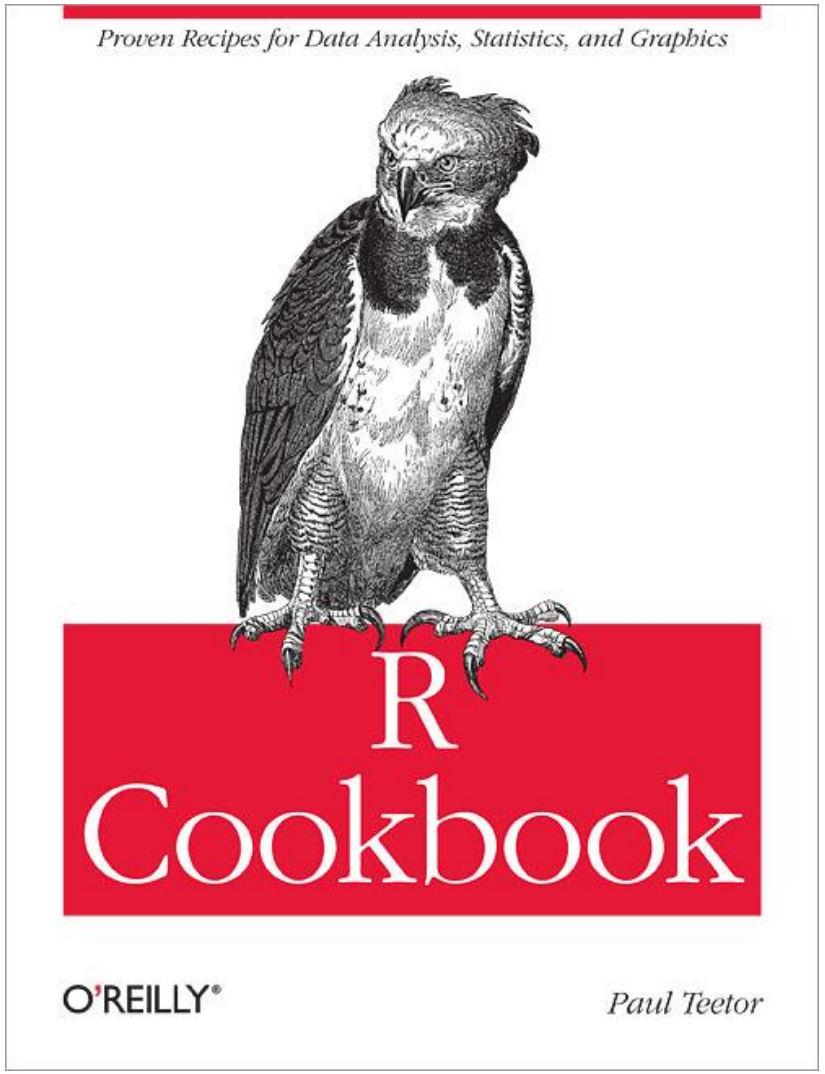


Bibliografía del curso

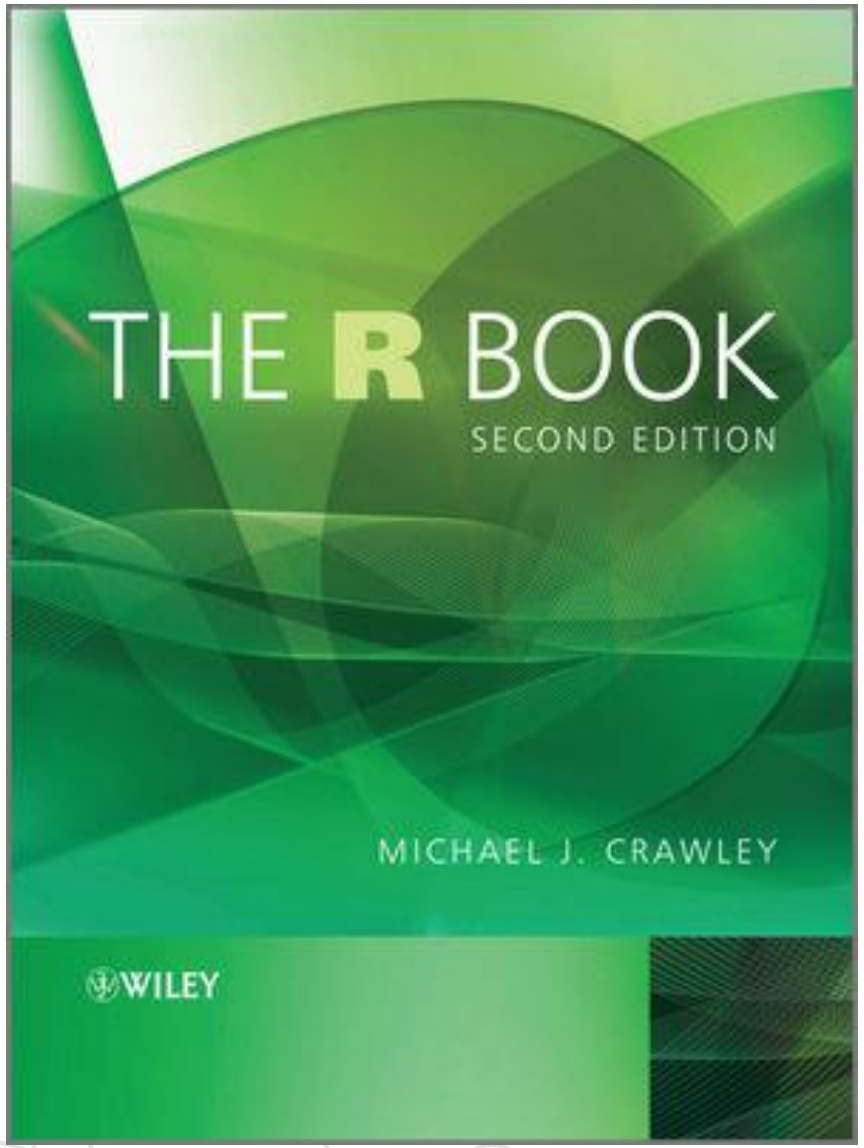


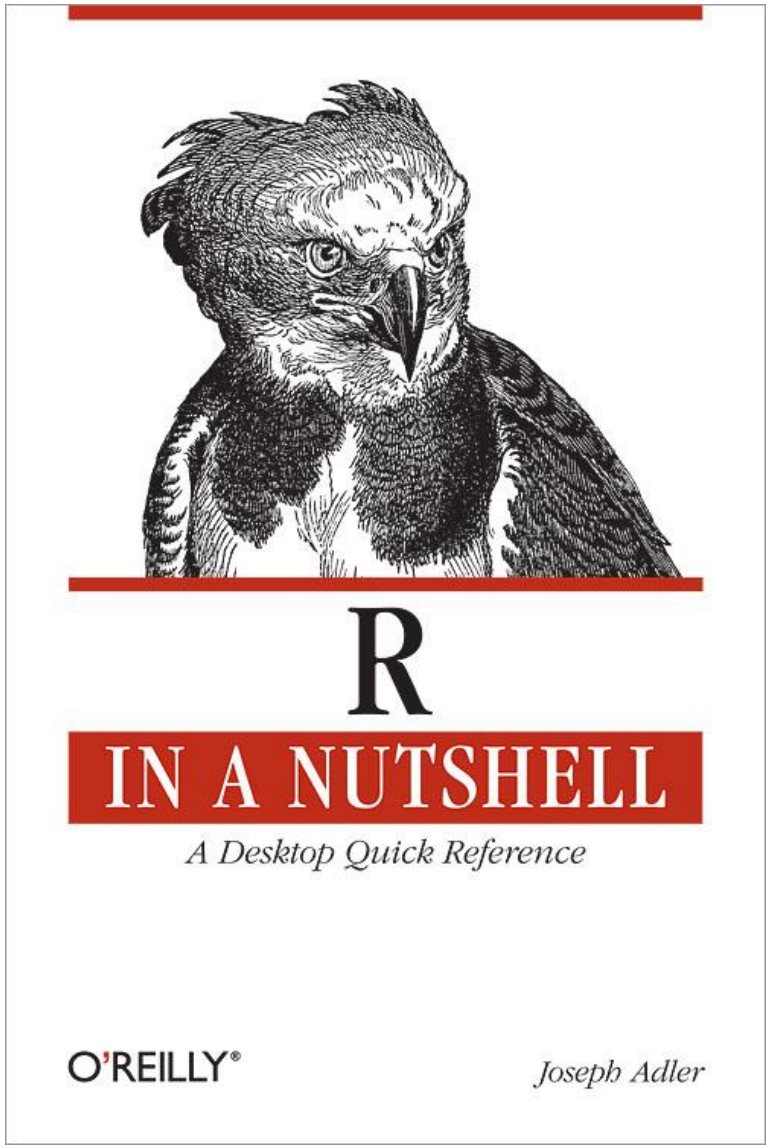
Bibliografía del curso



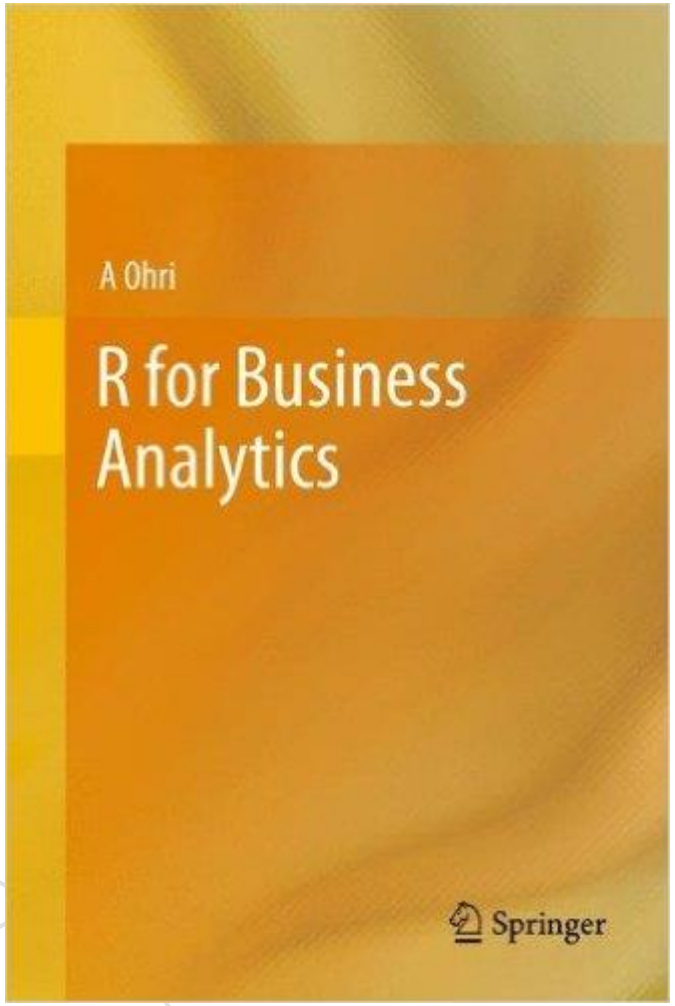


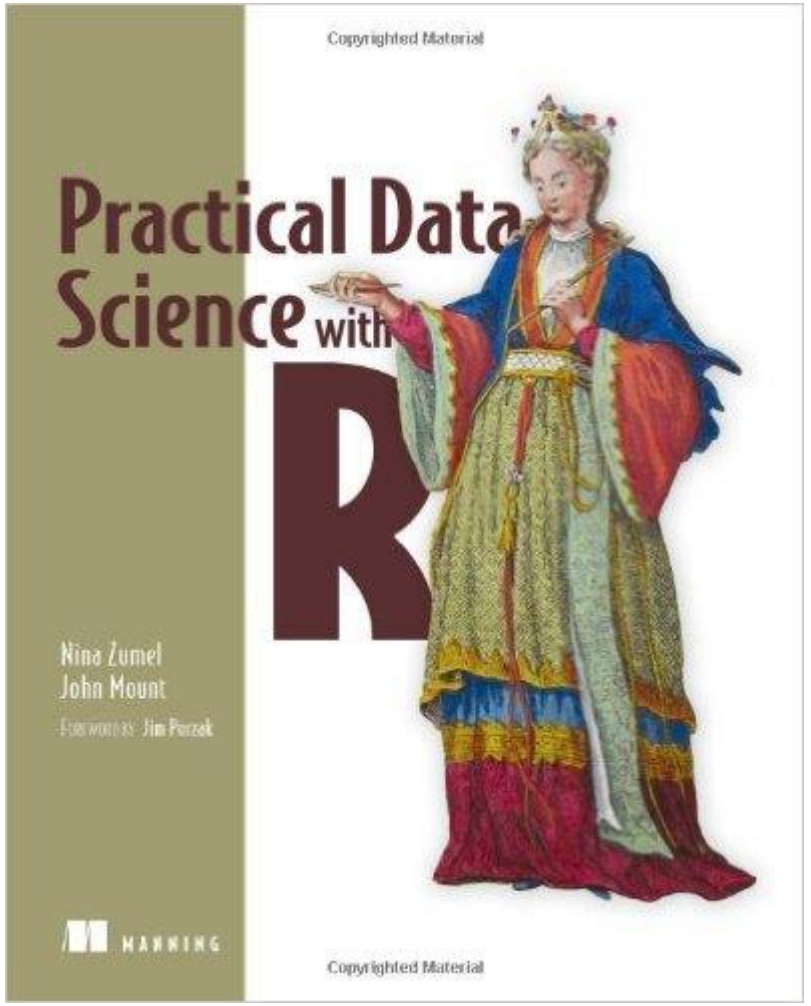
Bibliografía del curso



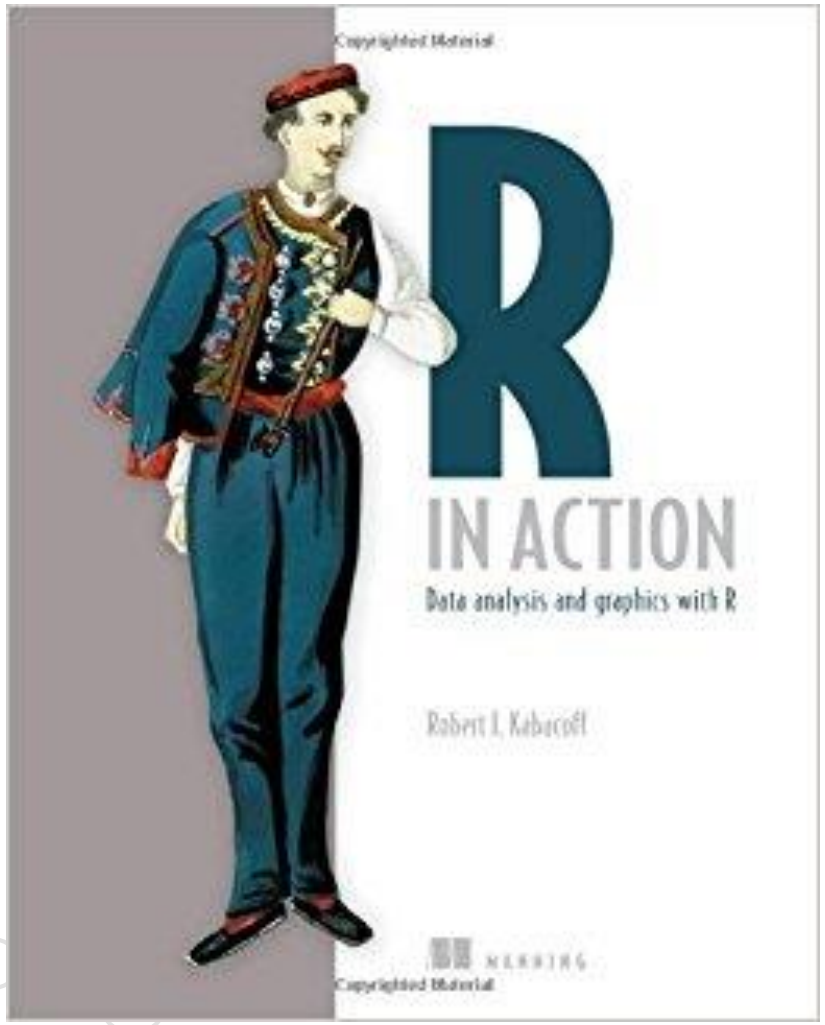


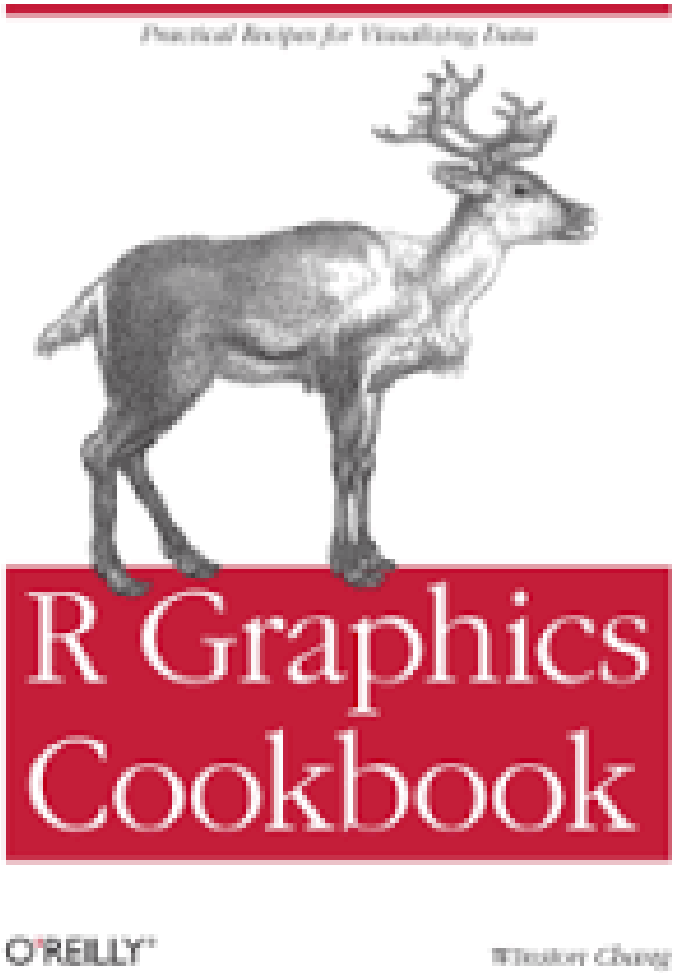
Bibliografía del curso



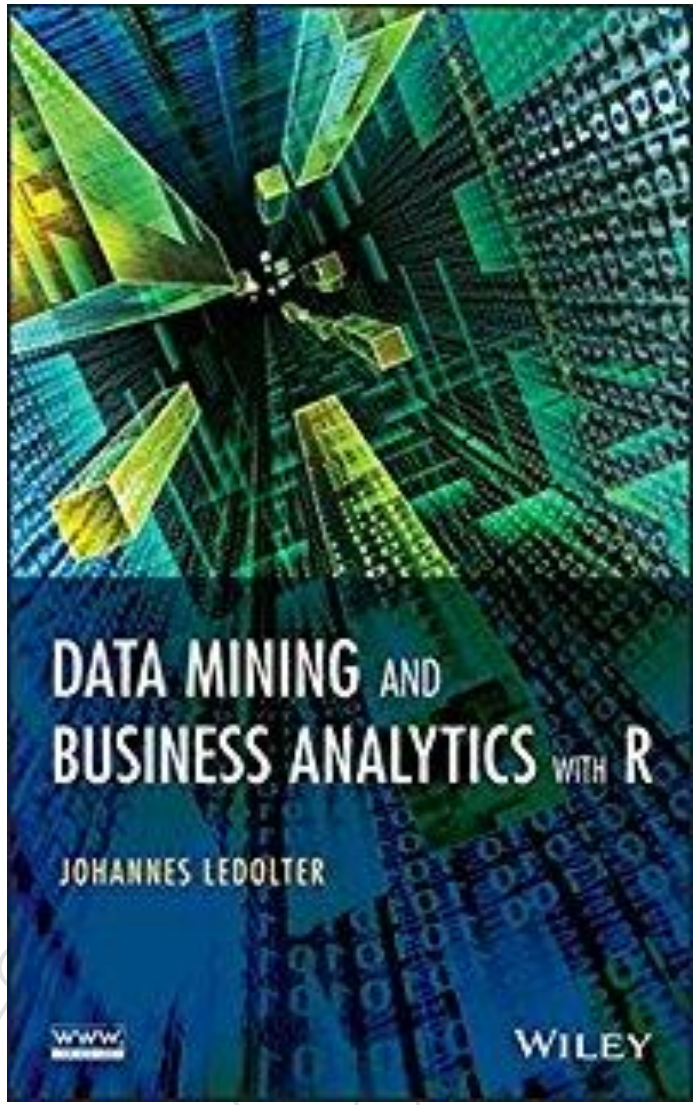


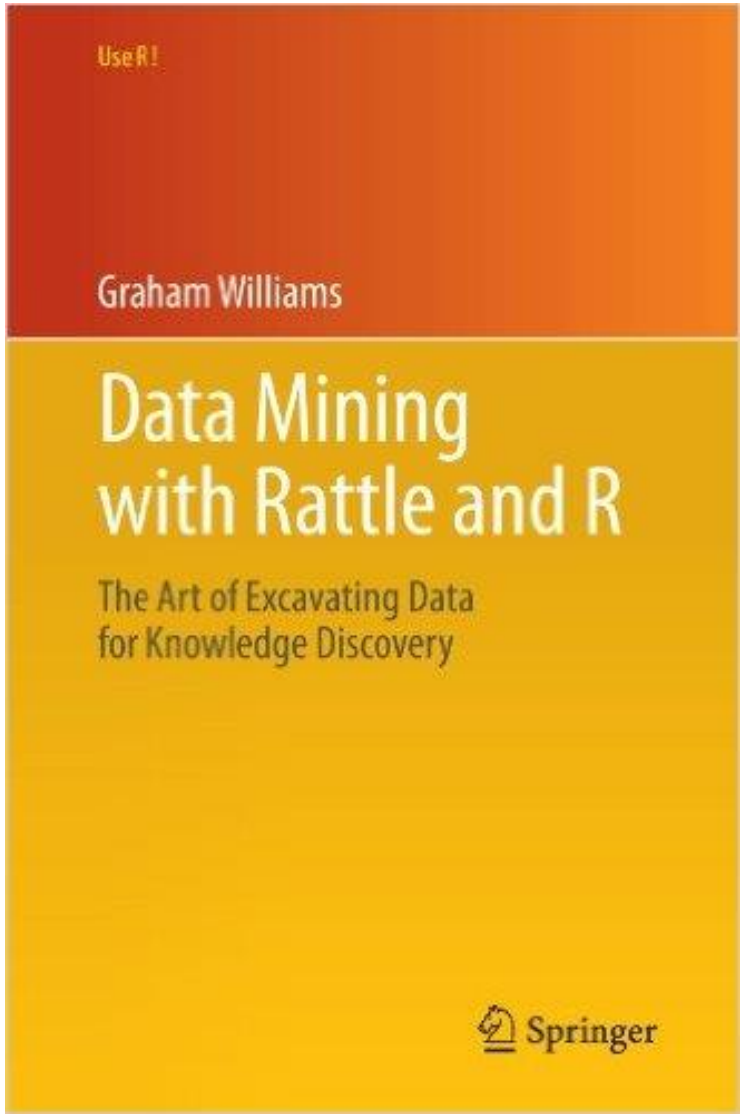
Bibliografía del curso





Bibliografía del curso





Data Mining Consulting

Bussiness Analytics

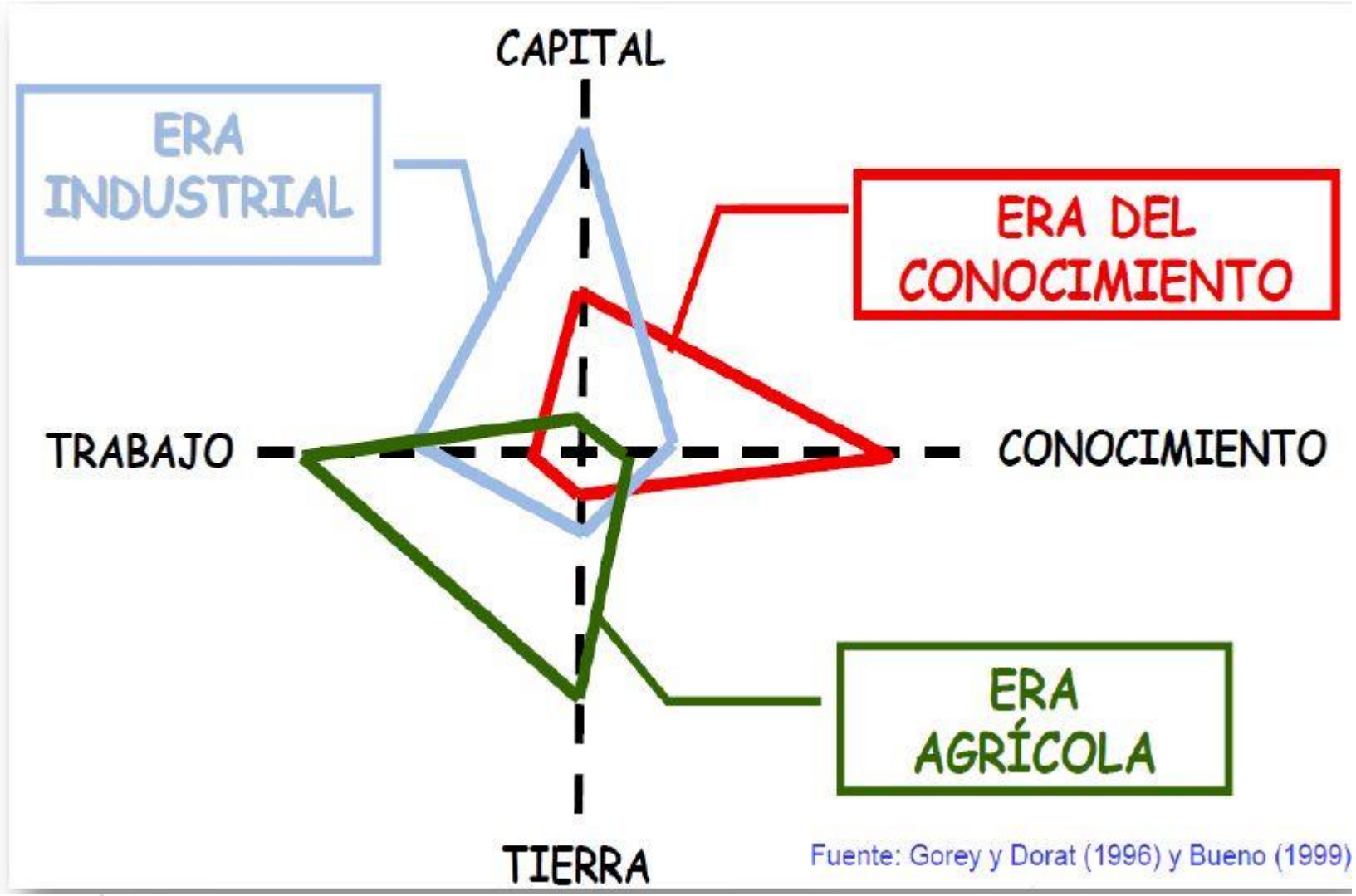
Mg Jesús Salinas Flores

jsalinas@lamolina.edu.pe

Data Mining Consulting

Necesidad de Conocimiento

Necesidad de conocimiento



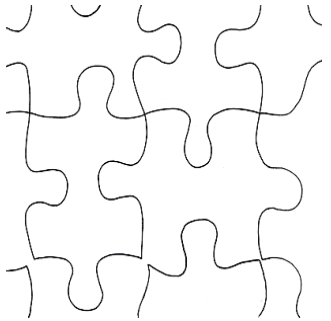
Fuente: Gorey y Dorat (1996) y Bueno (1999)

Descubrimiento de conocimiento

Problema de Negocio

Datos

Información (Resultados)



Conocimiento (Modelo)

Decisiones & Acciones

Business Analytics

El *Business Analytics* tiene un marcado enfoque en el análisis de la situación actual y la *predicción* de eventos futuros para entender el camino que tomará la empresa.

- Fuente: http://www.esan.edu.pe/apuntes-empresariales/2015/10/business-intelligence-vs-business-analytics-hay-diferencias/

Business Intelligence y Business Analytics

El *Business Intelligence* permite echar un vistazo al pasado de la empresa a través de análisis y reportes que tienen como base información histórica del negocio. Es ideal para comprender el *panorama de desarrollo histórico* de una empresa.

Por otro lado, el *Business Analytics* se enfoca en el análisis a *futuro* con base en la información de la empresa y modelos predictivos para apoyar la toma de decisiones y mejorar la competitividad del negocio.

- Fuente: http://www.esan.edu.pe/apuntes-empresariales/2015/10/business-intelligence-vs-business-analytics-hay-diferencias/

Data Mining Consulting

Business Analytics

Business Intelligence

"hace un uso extensivo de los datos, análisis estadístico, modelos explicativos y predictivos para impulsar la toma de decisiones"

"la generación de los datos e información para apoyar el mismo proceso."
"las aplicaciones y tecnologías para la recolección, almacenamiento, análisis y acceso a los datos para ayudar a una empresa a tomar mejores decisiones de negocios"

Diferencias entre BI y BA

	Business Intelligence	Business Analytics
Responde las preguntas…	¿Qué sucedió ? ¿Cuándo ? ¿Quién ? ¿Cuántos ?	¿Por qué sucedió ? ¿Ocurrirá otra vez ? ¿Qué pasaría si cambiamos X ? ¿Qué otras cosas dicen los datos, que nunca se nos ocurrió preguntar?
Incluye….	Reporting (KPIs, métricas) Sistemas de Monitoreo/Alertas Cuadros de Mando o Dashboards Scorecards OLAP (Cubos, Slice & Dice, Drilling) Consultas Adhoc / Querys	Análisis Estadístico / Cuantitativo Data Mining Modelamiento Predictivo Análisis Multivariado

Fuente: http://www.webmining.cl/2012/03/business-analytics-versus-business-intelligence/

Data Mining Consulting

Introducción al R

Mg Jesús Salinas Flores

jsalinas@lamolina.edu.pe

El entorno R

R es un lenguaje de programación para:

1. Manipulación de datos
2. Cálculo
3. Gráficos

Historia del R

Fue desarrollado inicialmente por ***R**obert Gentleman* y ***R**oss Ihaka* del Departamento de Estadística de la Universidad de Auckland en 1992. Sin embargo, inició en los Bell Laboratories de AT&T y ahora Alcatel-Lucent en Nueva Jersey con el lenguaje S.

El **lenguaje S** es un sistema para el análisis de datos desarrollado por John Chambers, Rick Becker, y colaboradores diferentes desde finales de 1970.

Historia del R

Los diseñadores iniciales, Gentleman y Ihaka, combinaron las fortalezas de dos lenguajes existentes, S y Scheme.

En sus propias palabras: "El lenguaje resultante es muy similar en apariencia a S, pero en el uso de fondo y la semántica es derivado desde Scheme". El resultado se llamó R "en parte al reconocimiento de la influencia de S y en parte para hacer gala de sus propios logros".

Su desarrollo actual es responsabilidad del *R Development Core Team*. contributors()

Fuente: https://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n)#Historia

Razones para aprender R

1. R es gratuito. No necesitan piratear o comprar licencias de software comercial.
2. R tiene una comunidad académica detrás que provee una muy buena documentación en línea.
3. R es empleado por investigadores de distintas áreas.
4. Constantemente aparecen nuevos paquetes gratuitos que expande la capacidad de R para solucionar diferentes problemas.
5. Emplea una interfase de línea de comando (command-line) de R que permite aprender mientras se hacen los cálculos

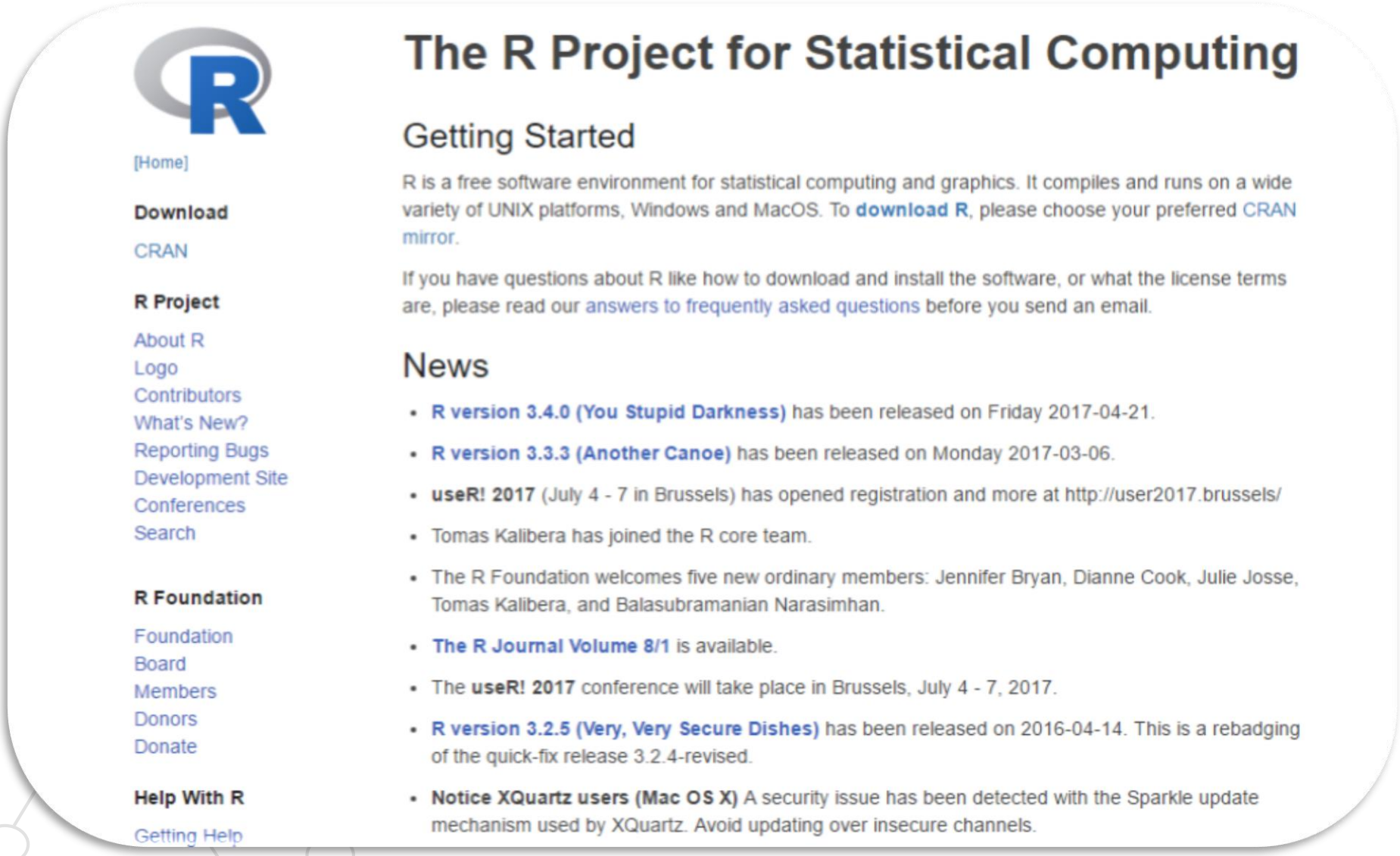
Razones para aprender R

6. R es uno de los paquetes estadísticos de mayor crecimiento en su uso en diferentes disciplinas.
7. El lenguaje de programación de R es intuitivo.
8. R crea gráficos de calidad superior a otros paquetes.
9. R y LaTex trabajan de manera integrada.
10. Es compatible con 'todos' los formatos de datos (csv, xls, sav, sas. . .)
11. R es multiplataforma, es decir funciona en Mac, Windows o Linux)

Fuente: http://www.icesi.edu.co/blogs/usando_r/2012/05/28/por-que-emplear-r-10-razones-por-las-que-un-estudiante-de-posgrado-deberia-usar-r/

Data Mining Consulting

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



[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

Help With R

Getting Help

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.

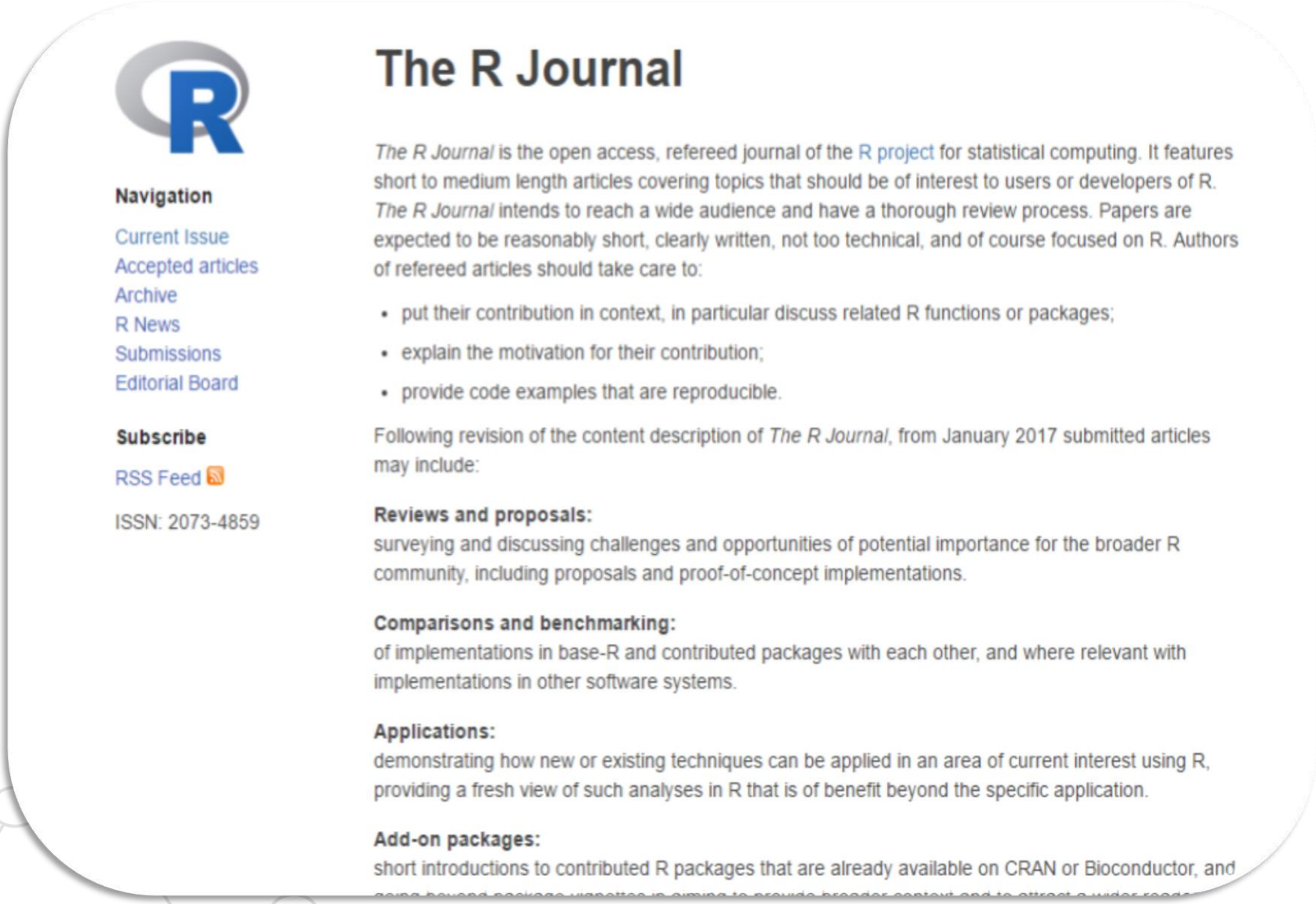
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 3.4.0 (You Stupid Darkness)** has been released on Friday 2017-04-21.
- **R version 3.3.3 (Another Canoe)** has been released on Monday 2017-03-06.
- **useR! 2017** (July 4 - 7 in Brussels) has opened registration and more at http://user2017.brussels/
- Tomas Kalibera has joined the R core team.
- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kalibera, and Balasubramanian Narasimhan.
- **The R Journal Volume 8/1** is available.
- The **useR! 2017** conference will take place in Brussels, July 4 - 7, 2017.
- **R version 3.2.5 (Very, Very Secure Dishes)** has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.
- **Notice XQuartz users (Mac OS X)** A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating over insecure channels.

Data Mining Consulting

https://journal.r-project.org/



The R Journal

The R Journal is the open access, refereed journal of the R project for statistical computing. It features short to medium length articles covering topics that should be of interest to users or developers of R. *The R Journal* intends to reach a wide audience and have a thorough review process. Papers are expected to be reasonably short, clearly written, not too technical, and of course focused on R. Authors of refereed articles should take care to:

- put their contribution in context, in particular discuss related R functions or packages;
- explain the motivation for their contribution;
- provide code examples that are reproducible.

Following revision of the content description of *The R Journal*, from January 2017 submitted articles may include:

Reviews and proposals:
surveying and discussing challenges and opportunities of potential importance for the broader R community, including proposals and proof-of-concept implementations.

Comparisons and benchmarking:
of implementations in base-R and contributed packages with each other, and where relevant with implementations in other software systems.

Applications:
demonstrating how new or existing techniques can be applied in an area of current interest using R, providing a fresh view of such analyses in R that is of benefit beyond the specific application.

Add-on packages:
short introductions to contributed R packages that are already available on CRAN or Bioconductor, and

Data Mining Consulting

https://www.r-bloggers.com

R-bloggers

R news and tutorials contributed by (750) R bloggers

Home | About | RSS | add your blog! | Learn R | R jobs | Contact us

WELCOME!

Follow @rbloggers 45.9K

Here you will find daily **news and tutorials about R**, contributed by over 750 bloggers.
There are many ways to **follow us** -
By e-mail:

Your e-mail here
Subscribe
41321 readers
BY FEEDBURNER

On Facebook:

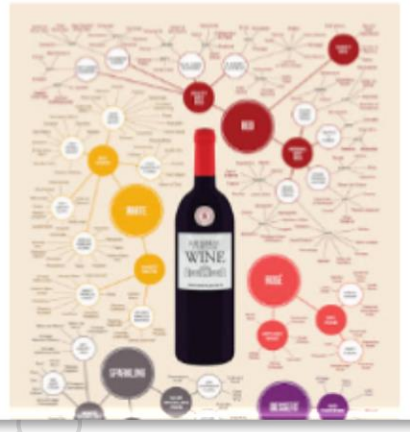
R blogg...
62.063 Me gusta

Te gusta

Machine Learning. Regression Trees and Model Trees (Predicting Wine Quality)

May 8, 2017

By Data Scientist PakinJa



We will develop a forecasting example using model trees and regression trees algorithms. The exercise was originally published in "Machine Learning in R" by Brett Lantz, PACKT publishing 2015 (open source community experience destilled).The example we will develop is about predicting...

Search & Hit Enter

RECENT POPULAR POSTS

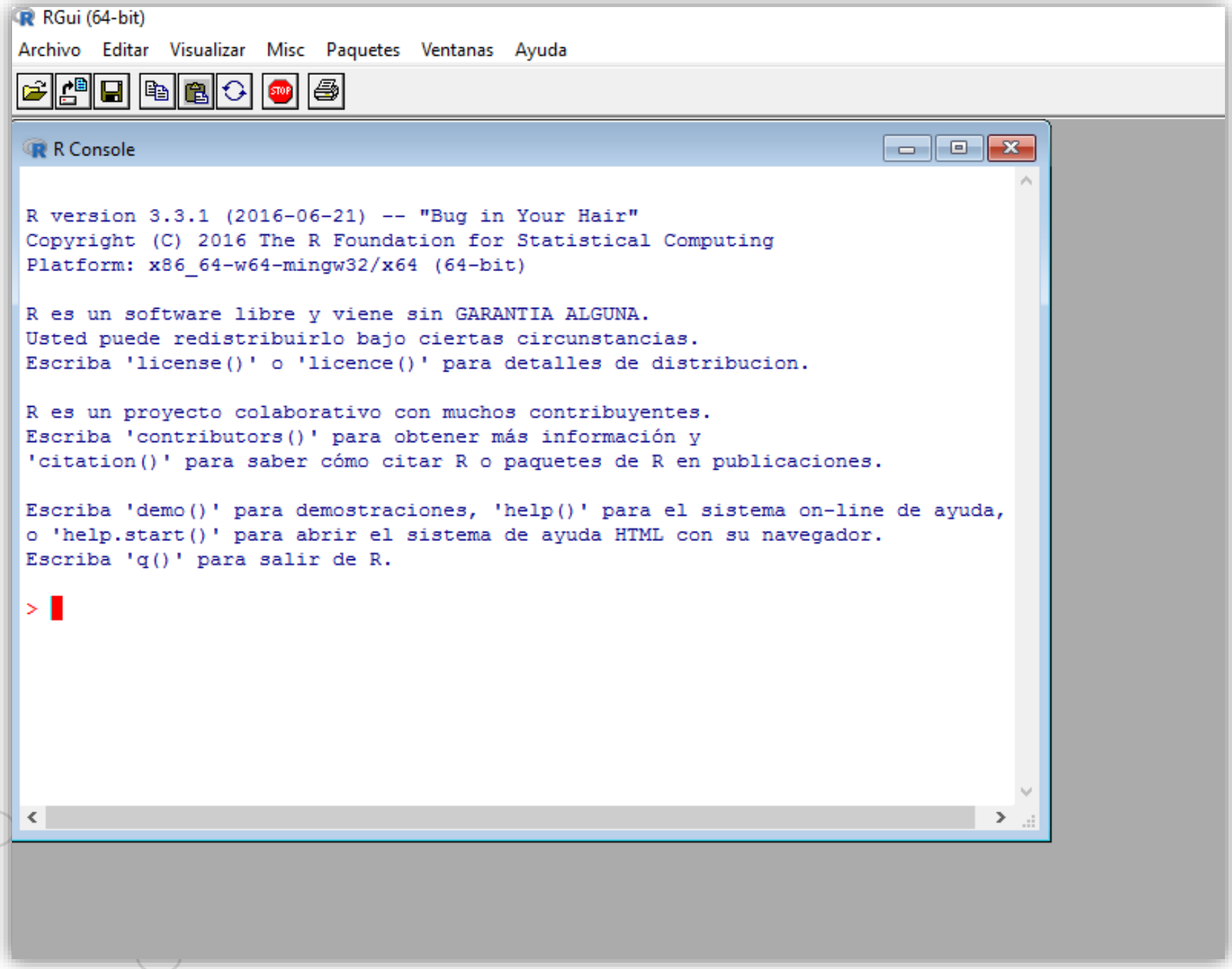
dplyr in Context
MIT Step by Step Instructions for Creating Your Own R Package.

MOST VISITED ARTICLES OF THE WEEK

1. How to write the first for loop in R
2. Installing R packages
3. How to perform a Logistic Regression in R
4. Using apply, sapply, lapply in R
5. How to Make a Histogram with Basic R
6. Tutorials for learning R
7. MIT Step by Step Instructions for Creating Your Own R Package.
8. Simple Linear Regression
9. dplyr in Context

Grupo en Facebook: R Project en español

Data Mining Consulting

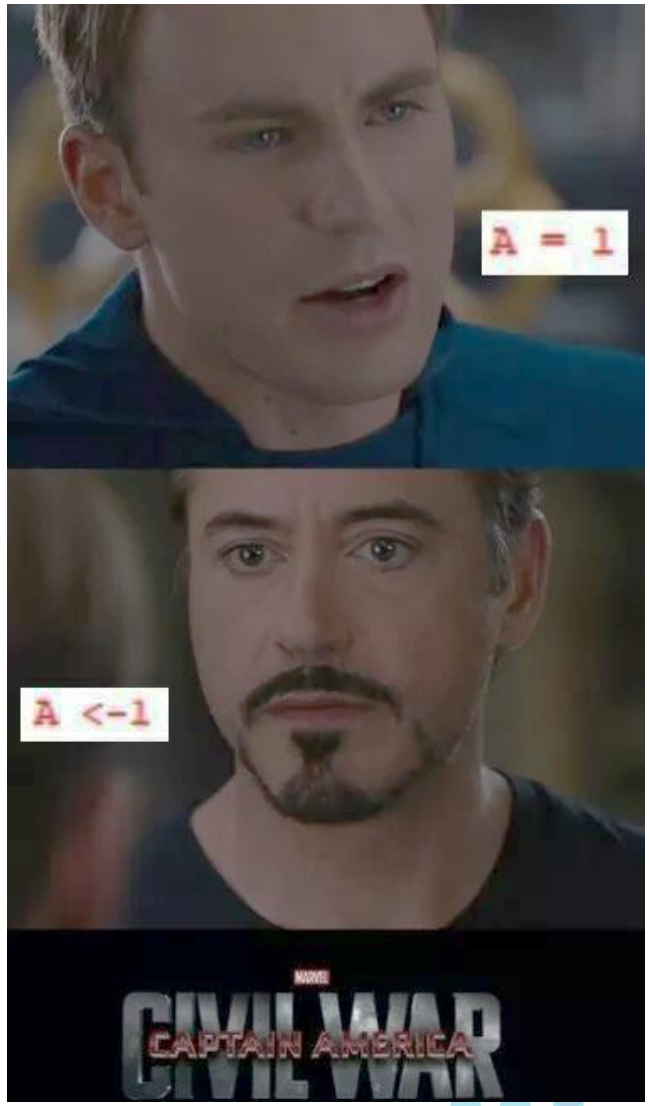


Data Mining Consulting

https://www.rstudio.com

Consideraciones con el R

- R distingue mayúsculas de minúsculas
- Las líneas de comentario comienzan con #
- El nombre de un objeto no debe empezar con un número
- Para asignar un contenido a un objeto se usa <- . En lugar de <- se puede usar =. Ejemplo x <- 10



Ayuda en R

¿Cómo obtener ayuda en R?

- help(), ?
- help.search(“”)
- help.start()

Objetos

¿Qué es un objeto en R?

- Las entidades que R crea y manipula se denominan objetos.
- Estos pueden ser: variables, cadenas de caracteres, funciones, etc.
- Un objeto puede ser creado con el símbolo <- o =

Área de trabajo (Workspace)

Todos los objetos en R se almacenan en el área de trabajo

- ls() Permite ver que objetos se encuentran en el workspace.
- save.image() Permite guardar el workspace
- load() Permite recuperar un workspace guardado
- rm() Permite eliminar objetos del workspace

Paquetes

Tres niveles de funciones

- Las que están accessible por defecto.
- Las que están instaladas pero no inmediatamente accesibles.
- Las que no estén instaladas pero están disponible en paquetes o packages.

Data Mining Consulting

https://cloud.r-project.org/

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

Contributed Packages

Available Packages

Currently, the CRAN package repository features 10577 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. The manual R Installation and Administration (also contained in the R base sources) explains the process in detail.

CRAN Task Views allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 35 views are available.

Package Check Results

All packages are tested regularly on machines running Debian GNU/Linux, Fedora, OS X, Solaris and Windows.

The results are summarized in the check summary (some timings are also available). Additional details for Windows checking and building can be found in the Windows check summary.

Writing Your Own Packages

The manual Writing R Extensions (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.

Repository Policies

The manual CRAN Repository Policy [PDF] describes the policies in place for the CRAN package repository.

Related Directories

Data Mining Consulting

http://cran.r-project.org/web/views

CRAN Task Views

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
ExtremeValue	Extreme Value Analysis
Finance	Empirical Finance
FunctionalData	Functional Data Analysis
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods

Scripts

¿Qué es un script en R?

- Un script es un archivo de texto que contiene los commandos que uno ingresa en la línea de commandos del R.
- Para escribir un script se puede usar cualquier editor de texto.
- Los scripts creados en R tienen extensión *.R

Data Mining Consulting

Caso de Estudio

Mg Jesús Salinas Flores

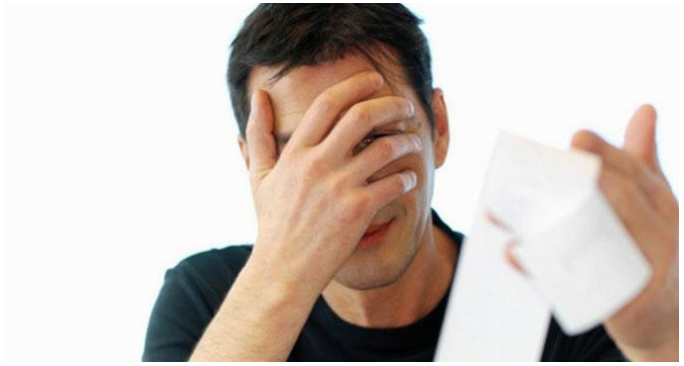
jesus.salinas@dataminingperu.com

Caso de estudio:
Identificación de morosos en un producto crediticio

La información a analizar proviene de un producto crediticio de una institución financiera.

El objetivo es predecir si un **nuevo cliente** de la institución financiera podría ser clasificado como moroso o no moroso.

Se considera recolectar información respecto a 11 atributos registrados de un cliente al momento que este se afilia a dicha institución financiera.



Data Mining Consulting

Riesgo_morosidad.csv

edad	sexo	nrodepen	fonopart	fonolab	autovaluo	esaval	tieneaval	antiguedad	tiporenta	dpto	morosidad
44	Femenino	3	Si	Si	No	No	No	63	Fijo	Lima	No Moroso
77	Femenino	4	Si	Si	No	No	Si	62	Fijo	Lima	Moroso
59	Femenino	5	No	No	No	Si	No	59	Fijo	Lima	Moroso
35	Femenino	5	No	Si	No	No	No	58	Fijo	Lima	Moroso
65	Femenino	0	Si	No	No	Si	No	56	Fijo	Lima	No Moroso
66	Masculino	1	Si	Si	Si	Si	Si	54	Fijo	Lima	No Moroso
73	Masculino	5	Si	No	No	No	No	53	Fijo	Lima	Moroso
73	Femenino	0	No	No	No	Si	Si	53	Fijo	Lima	No Moroso
74	Femenino	0	Si	No	No	Si	No	53	Fijo	Lima	No Moroso
75	Femenino	1	No	No	No	No	No	53	Fijo	Lima	Moroso
52	Masculino	0	Si	No	No	No	No	52	Variable	Lima	No Moroso
76	Masculino	0	Si	Si	Si	No	Si	52	Fijo	Lima	No Moroso

Variables Independientes / Predictoras

Variable Dependiente / a Predecir

Riesgo_morosidad.csv

edad	sexo	nrodepen	fonopart	fonolab	autovaluo	esaval	tieneaval	antiguedad	tiporenta	dpto	morosidad	clase.pred	proba.pred
44	Femenino	3	Si	Si	No	No	No	63	Fijo	Lima	No Moroso	No Moroso	0.04227578
77	Femenino	4	Si	Si	No	No	Si	62	Fijo	Lima	Moroso	Moroso	0.96886035
59	Femenino	5	No	No	No	Si	No	59	Fijo	Lima	Moroso	Moroso	0.96886035
35	Femenino	5	No	Si	No	No	No	58	Fijo	Lima	Moroso	Moroso	0.96886035
65	Femenino	0	Si	No	No	Si	No	56	Fijo	Lima	No Moroso	No Moroso	0.04227578
66	Masculino	1	Si	Si	Si	Si	Si	54	Fijo	Lima	No Moroso	No Moroso	0.04227578
73	Masculino	5	Si	No	No	No	No	53	Fijo	Lima	Moroso	Moroso	0.96886035
73	Femenino	0	No	No	No	Si	Si	53	Fijo	Lima	No Moroso	No Moroso	0.04227578
74	Femenino	0	Si	No	No	Si	No	53	Fijo	Lima	No Moroso	No Moroso	0.04227578
75	Femenino	1	No	No	No	No	No	53	Fijo	Lima	Moroso	No Moroso	0.04227578
52	Masculino	0	Si	No	No	No	No	52	Variable	Lima	No Moroso	No Moroso	0.04227578
76	Masculino	0	Si	Si	Si	No	Si	52	Fijo	Lima	No Moroso	No Moroso	0.04227578

Variables Independientes / Predictoras

Variable Dependiente / a Predecir

Variables del modelo