Introduction to
Data Science and Analytics

Stephan Sorger
www.StephanSorger.com

Unit 7. R Essentials

Outline/ Learning Objectives

Topic	Description
Introduction	Analytics and statistical software
Suppliers	Major suppliers of statistical analytics software
Functions	Basic functions and features of R
Session	Sample working session in R; Linear regression
Resources	Where to learn more about R

Analytics and Statistical Analysis Software: Introduction

Topic	Definition
Definition	Software designed for in-depth analysis Unlike MS Excel (general purpose spreadsheet)
Origins	SAS conceived in 1966 by Anthony J. Barr Placed statistical procedures in formatted file framework
Uses	Advanced statistical techniques Nonlinear functions; Multiple regression; Conjoint
Advantages	Powerful; Accurate; Specific tools
Disadvantages	Command line interface; steep learning curve Very expensive

Analytics and Statistical Analysis Software: Major Suppliers

Criteria	SAS	SPSS	R
Market	Fortune 500	Universities	Universities
Focus	Power	Ease of use	Price
User	Power user	Student	Price-sensitive
Origins	Industry	Education	Open Source
Learning	Difficult	Moderate	Moderate
Cost	$86,600/yr+	$16,000/yr+	Free
UI	Command Line	Point & Click	Command Line
Database	32,768 var.	1 file at a time	
Graphics	SAS/Graph	High quality	Different packages
Analogy	Microsoft	Apple	Linux

UCLA, Statistical Software Packages Comparison, ats.ucla.edu:
http://www.ats.ucla.edu/stat/mult_pkg/compare_packages.htm

MineQuest Business Analytics, “Cost of Licensing WPS 3.0 vs. SAS 9.3.” February 2013.
http://www.minequest.com/downloads/Pricing_Comparisons_Between_WPS_and_SAS.pdf

IBM SPSS Statistics website, “Buy IBM SPSS Statistics Now”
http://www-01.ibm.com/software/analytics/spss/products/statistics/buy-now.html

R: Introduction

Topic	Description
Description	Free statistical computing and graphics software package Widely used among statisticians and data miners Increased popularity in 2010 - on
History	Started in 1993 as implementation of S programming language (1976) R developed by Ross Ihaka and Robert Gentleman “R” from Ross & Robert, as well as play on “S”
Functions	R includes many functions, which can be expanded through packages
Data	Can handle multiple simultaneous data sets, unlike Excel Data types: scalars, vectors, matrices, data frames, and lists Vectors: numerical, character, logical
Commercial	Revolution Analytics offers enterprise version ($); Purchased by Microsoft

References:
1. Venables, W.N., Smith, D.M., “An Introduction to R.” Version 3.0.1. May 16, 2013.
http://www.cran.r-project.org/doc/manuals/R-intro.pdf

R: Essentials

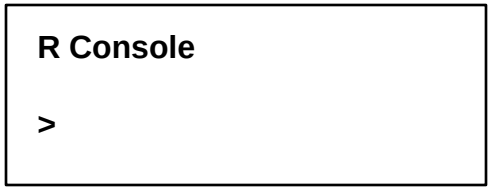
Topic	Description
Commands	Based on UNIX; case sensitive Commands separated by “;” or by newline <CR>
Comments	#Hashtags to indicate comments
Prompt	> #system is waiting for you to type something Traditional version not menu-driven, unlike consumer software
Arithmetic	> 5 + 4 [1] 9 #system returns the sum of 5 + 4, which is 9
Assignment (=)	> x <- 3 # assign the number “3” to the object “x”; similar to “=“ sign
Help	2 ways to get help; Example: Get help with “read.csv” command ?(read.csv) help(read.csv)

R: Essentials

Topic	Description
Functions	R features a rich set of functions c() : Function c Statistics functions: mean(x); median(x); range(x); etc. Arithmetic functions: 4^2; log (10); sqrt (16)
Vector	> x <- c(1, 2, 3) # assign a vector of numbers to the object x
Matrix	> y <- matrix(c(1, 2, 3, 4, 5, 6), 2, 3 # create 2 x 3 matrix
Print	Ask R to print out numbers inside an object, such as a vector by printing it > print (x) # ask R to print out x > x # Or, you can just type the variable and hit return
Plot	Ask R to plot out lines based on a dataset by plotting the data > plot(data)
Small subset	R is a large, complex language. We cover only a small % in this class https://cran.r-project.org/doc/contrib/Short-refcard.pdf

R: Getting Started

Topic	Description
Download R	Windows: http://cran.r-project.org/bin/windows/base/ Mac: http://cran.r-project.org/bin/macosx/
Launch R	Double-click to launch Will see prompt in “R Console” >



Sample R Session: Regression Analysis

Topic	Description
1. Preparation	Remove introductory content; First line should be data headers Save Excel file as Comma Separated Values (CSV)
2. Directory	Optional: Set up working directory for dataset; allows shorter filepaths Windows: See “Windows Explorer help” for more info Mac: See “Finder help” for more info
3. Filename	Need complete filename Example: “C:\My Documents\Folder A\Filename.csv” Alternative 1: Right click to see filename Alternative 2: Find filename in Windows Explorer (Windows); Finder (Mac) Alternative 3: Drag csv file and drop into R Console; Will show filename

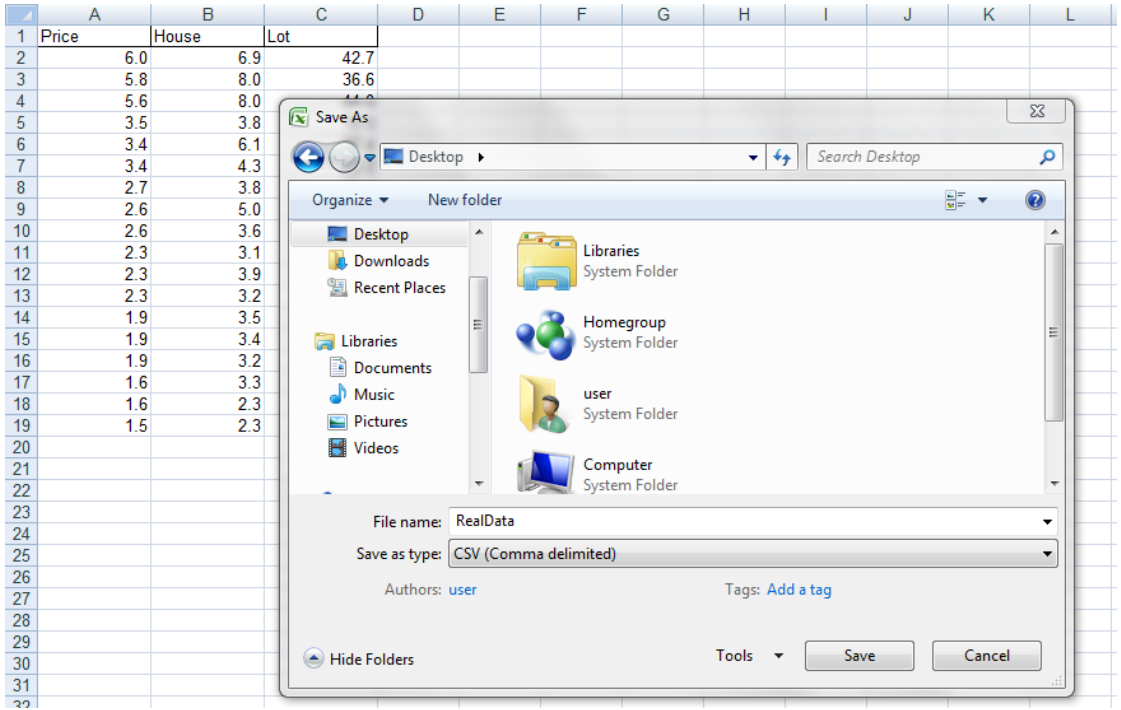
Sample R Session: Regression Analysis

Topic	Description
4. Read CSV data	Datafile <- read.csv("C:\\My Documents\\Desktop\\Filename.csv", header=T)
5. Check data	Print out dataset to ensure it was loaded correctly print(Datafile): will print out entire datafile; OK for small datasets str(Datafile): Shows structure of Datafile; "data.frame: 4 obs. of 4 variables" summary(Datafile): Shows summary: Min; Max; Mean; Median
6. Run regression	lm: Regression analysis in R; stands for Linear Model lm(Dependent~Independent+Independent, Dataset)
7. Interpret Results	Compare results obtained with R with those from Microsoft Excel

R Example: Causal Analysis Forecast for Real Estate

Step	Description
1. Preparation	Remove introductory information; First row = header row

“Save As” CSV



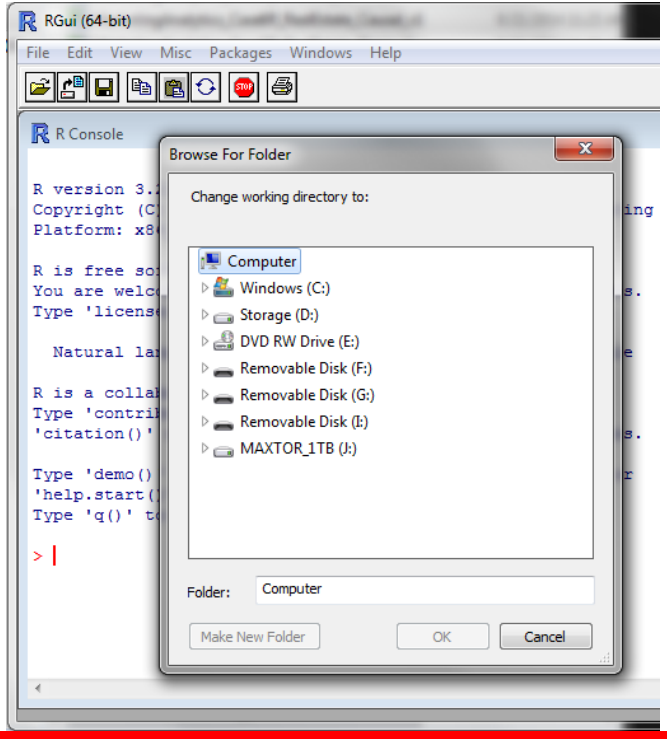
	A	B	C
1	Price	House	Lot
2	6.0	6.9	42.7
3	5.8	8.0	36.6
4	5.6	8.0	
5	3.5	3.8	
6	3.4	6.1	
7	3.4	4.3	
8	2.7	3.8	
9	2.6	5.0	
10	2.6	3.6	
11	2.3	3.1	
12	2.3	3.9	
13	2.3	3.2	
14	1.9	3.5	
15	1.9	3.4	
16	1.9	3.2	
17	1.6	3.3	
18	1.6	2.3	
19	1.5	2.3	

R Example: Causal Analysis Forecast for Real Estate

Step	Description
2. Directory	Optional: Can set up working directory

In R, select
File → Change dir…

then select where you want to put R files

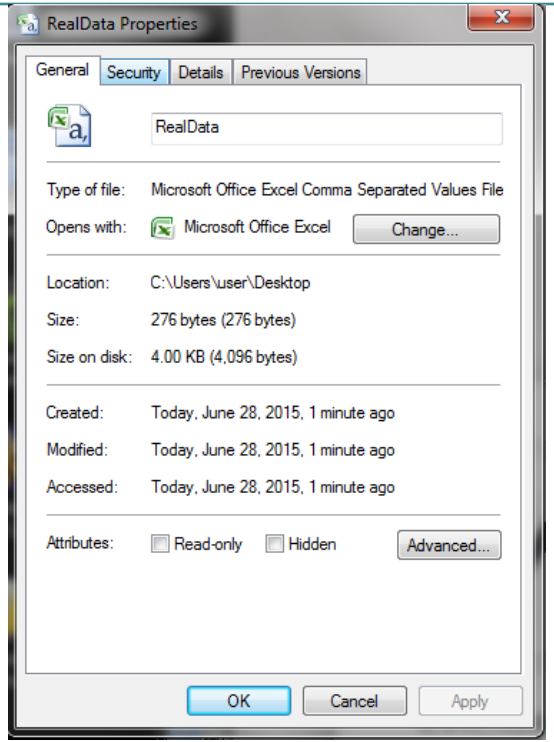


R Example: Causal Analysis Forecast for Real Estate

Step	Description
3. Filename	“C:\\Users\\user\\Desktop\\RealData.csv”

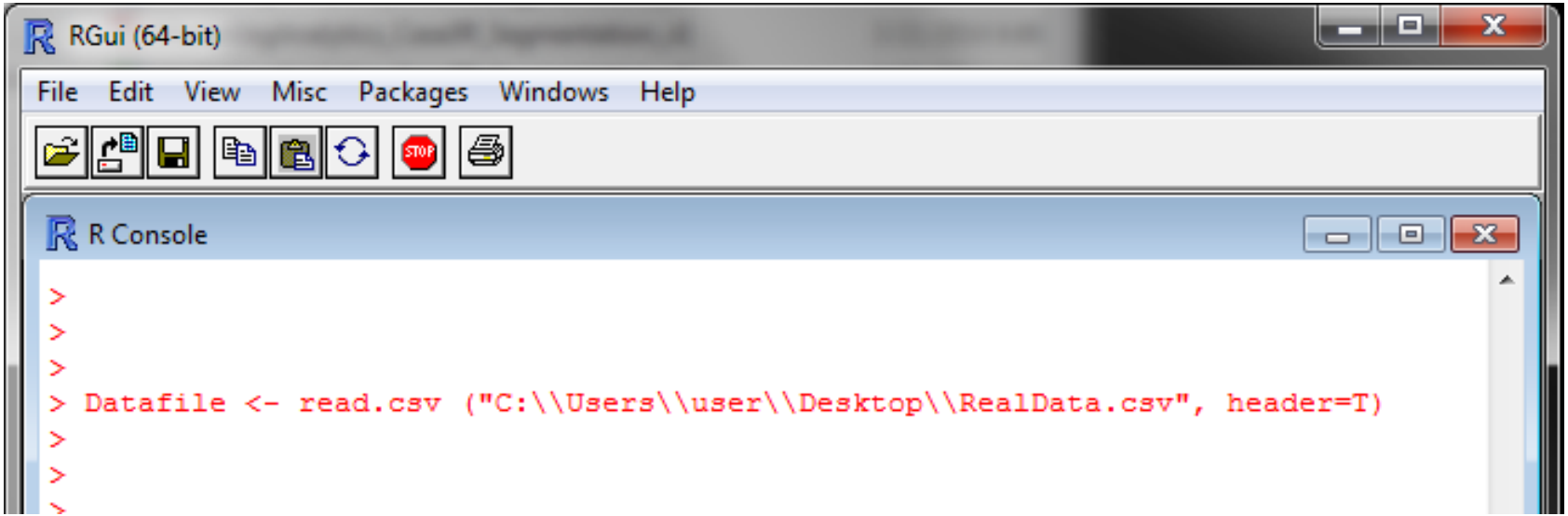
Windows:
Right-click on file
to get file properties;
will show full filename
under “Location”

Mac:
Check Finder
to find full filename
OR:
Drag file into R



R Example: Causal Analysis Forecast for Real Estate

Step	Description
4. Read Data	Datafile <- read.csv("C:\\Users\\user\\Desktop\\RealData.csv", header=T)



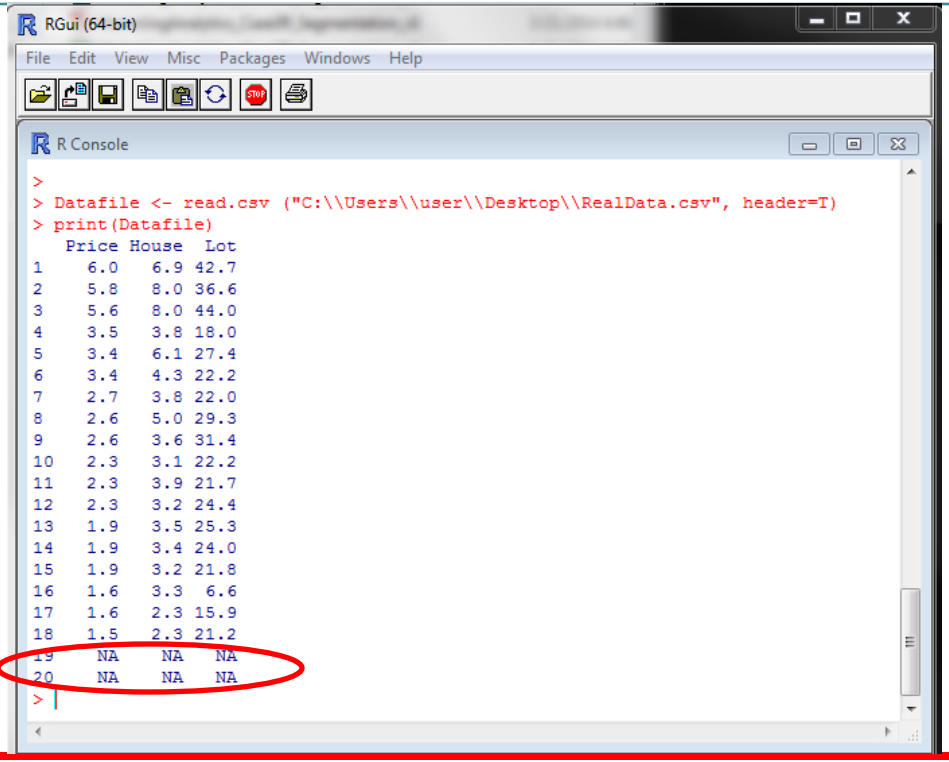
Alternative: Set up working directory

R Example: Causal Analysis Forecast for Real Estate

Step	Description
5. Check Data	print (Datafile) ; check if dataset looks OK

For large datasets, ask R to provide summary data instead of printing out entire dataset

Looks good, but we should substitute "0" values for "NA"



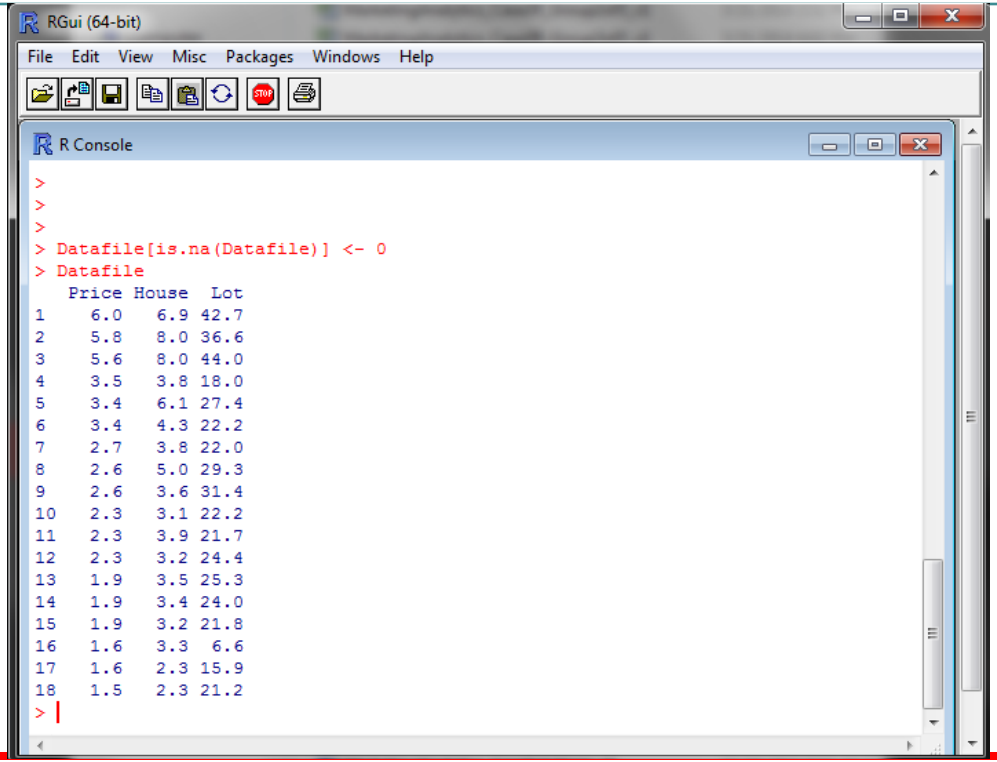
R Example: Causal Analysis Forecast for Real Estate

Step	Description
5. Check Data	print (Datafile) ; check if dataset looks OK

To substitute “0” for NA,
use the “is.na() function:

Datafile [is.na (Datafile)] <- 0

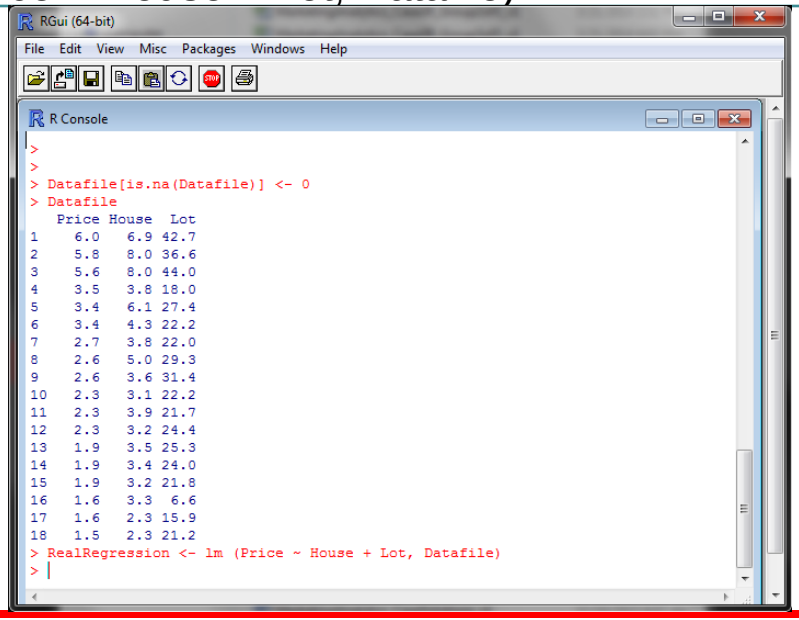
NA’s are gone!



R Example: Causal Analysis Forecast for Real Estate

Step	Description
6. Run Regression	lm(Dependent~Independent+Independent, Dataset) Dependent variable: Price; Independent variable: House; Lot Equation: Price = c1 + c2*(House Size) + c3*(Lot Size) RealRegression <- lm(Price ~ House + Lot, Datafile)

Find tilde symbol “ ~ “
at upper left of keyboard,
to left of number “1”

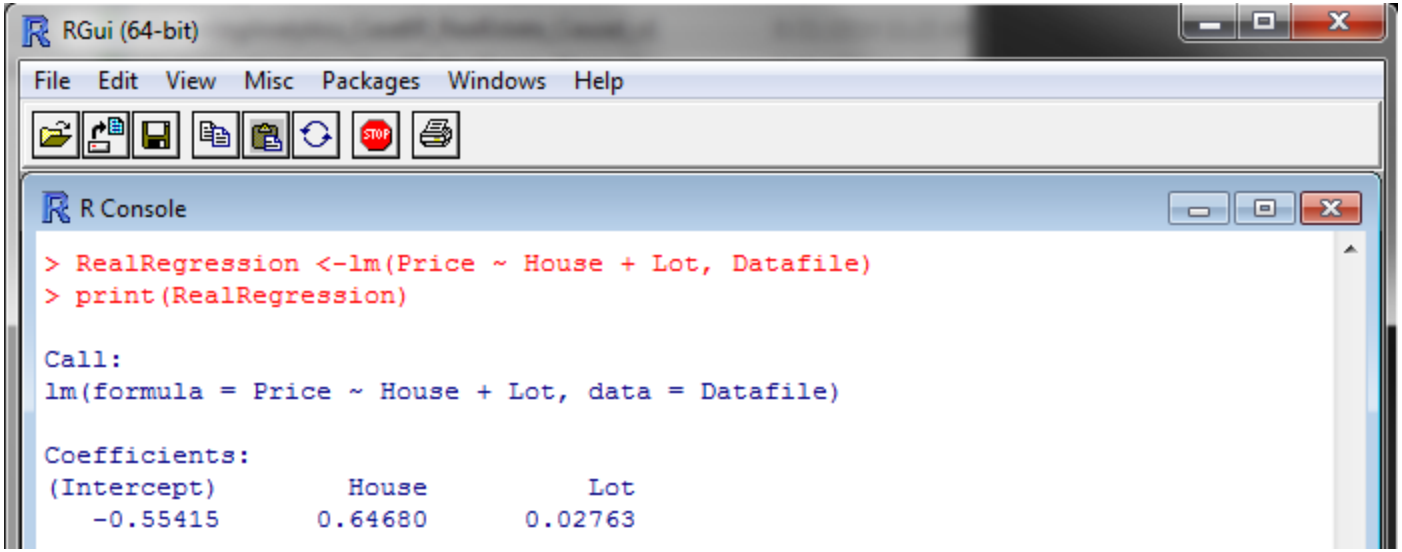


R Example: Causal Analysis Forecast for Real Estate

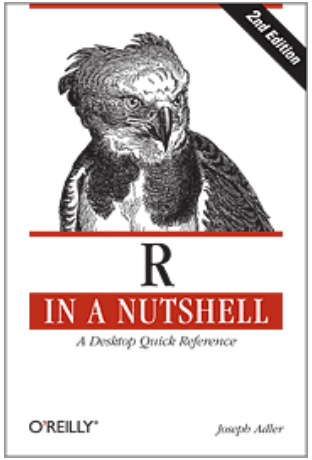
Topic	Description
7. Interpret Results	Compare results from R with those from Excel

Method	Coefficient	House Size	Lot Size
Excel	-0.554	+0.646	+0.027
R	-0.55415	+0.64680	+0.02763

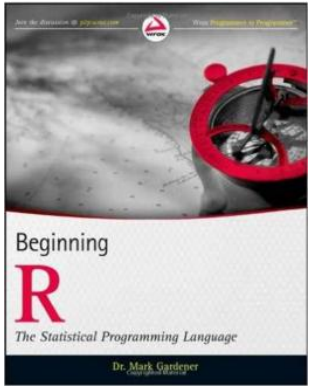
R results agree well with those of Excel



R Resources: Learning More About R: Print Books

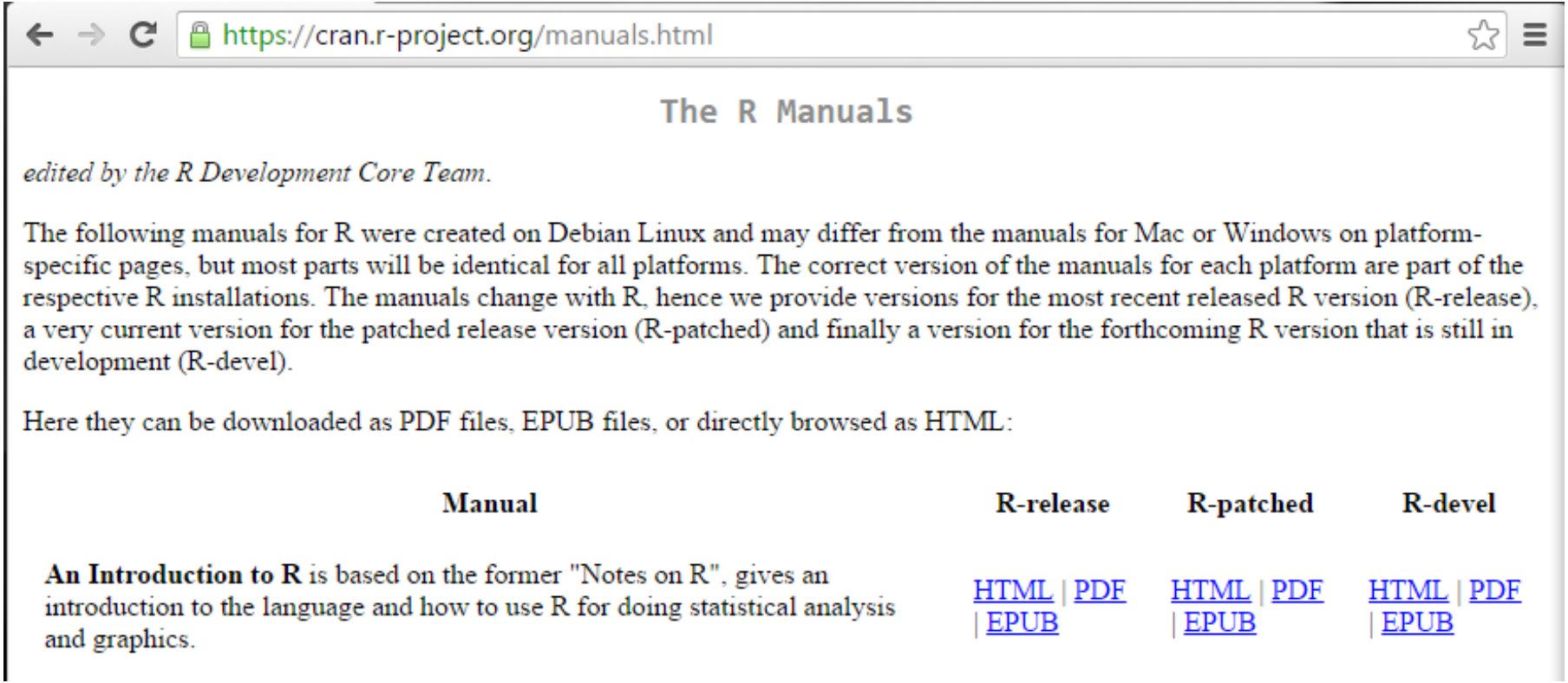


R in a Nutshell
By Joseph Adler
Published by O'Reilly Media



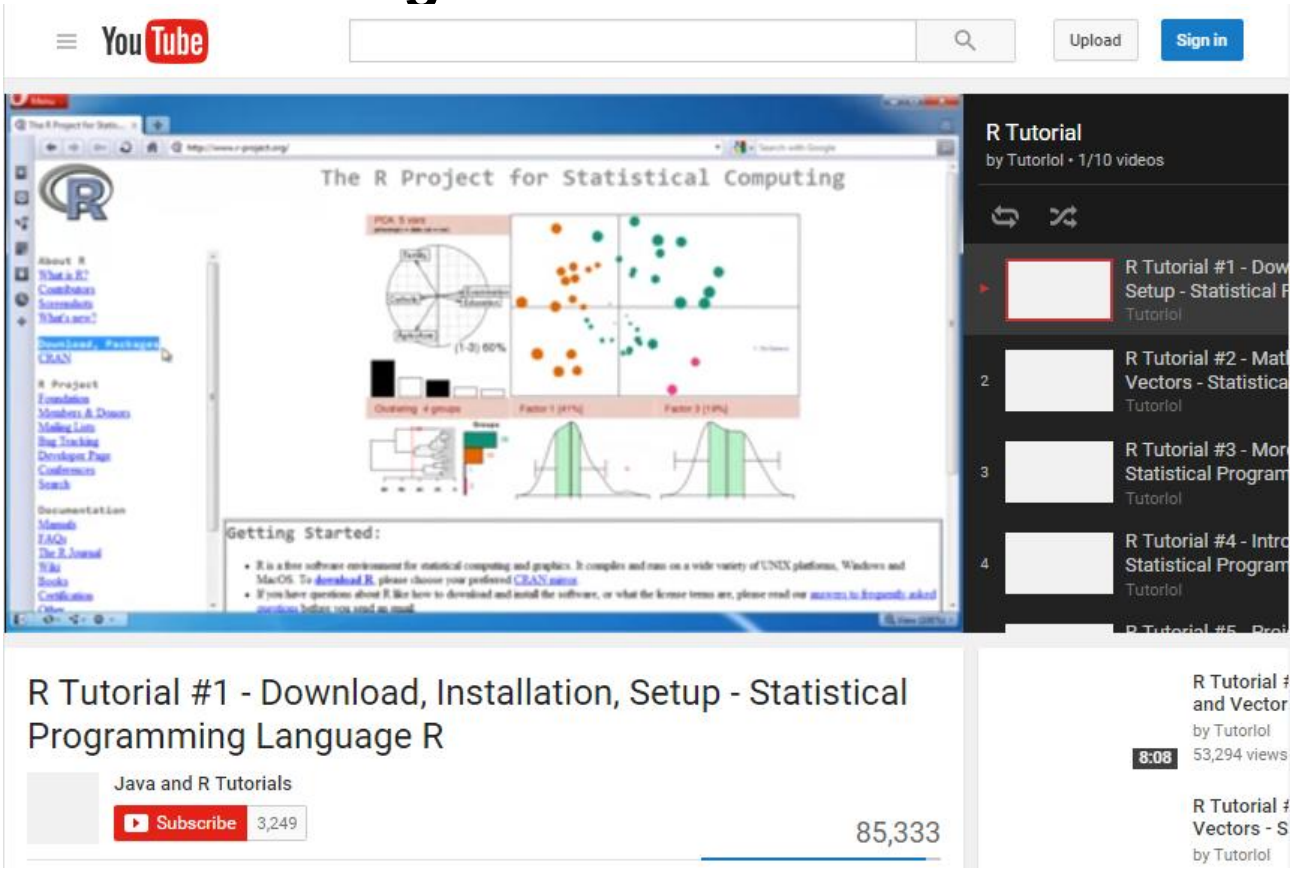
Beginning R: The Statistical Programming Language
By Mark Gardener
Published by John Wiley & Sons

R Resources: Learning More About R: Online Text



https://cran.r-project.org/manuals.html

R Resources: Learning More About R: YouTube



https://www.youtube.com/watch?v=ZoPJGmpYJzw

Outline/ Learning Objectives

Topic	Description
Introduction	Analytics and statistical software
Suppliers	Major suppliers of statistical analytics software
Functions	Basic functions and features of R
Session	Sample working session in R; Linear regression
Resources	Where to learn more about R