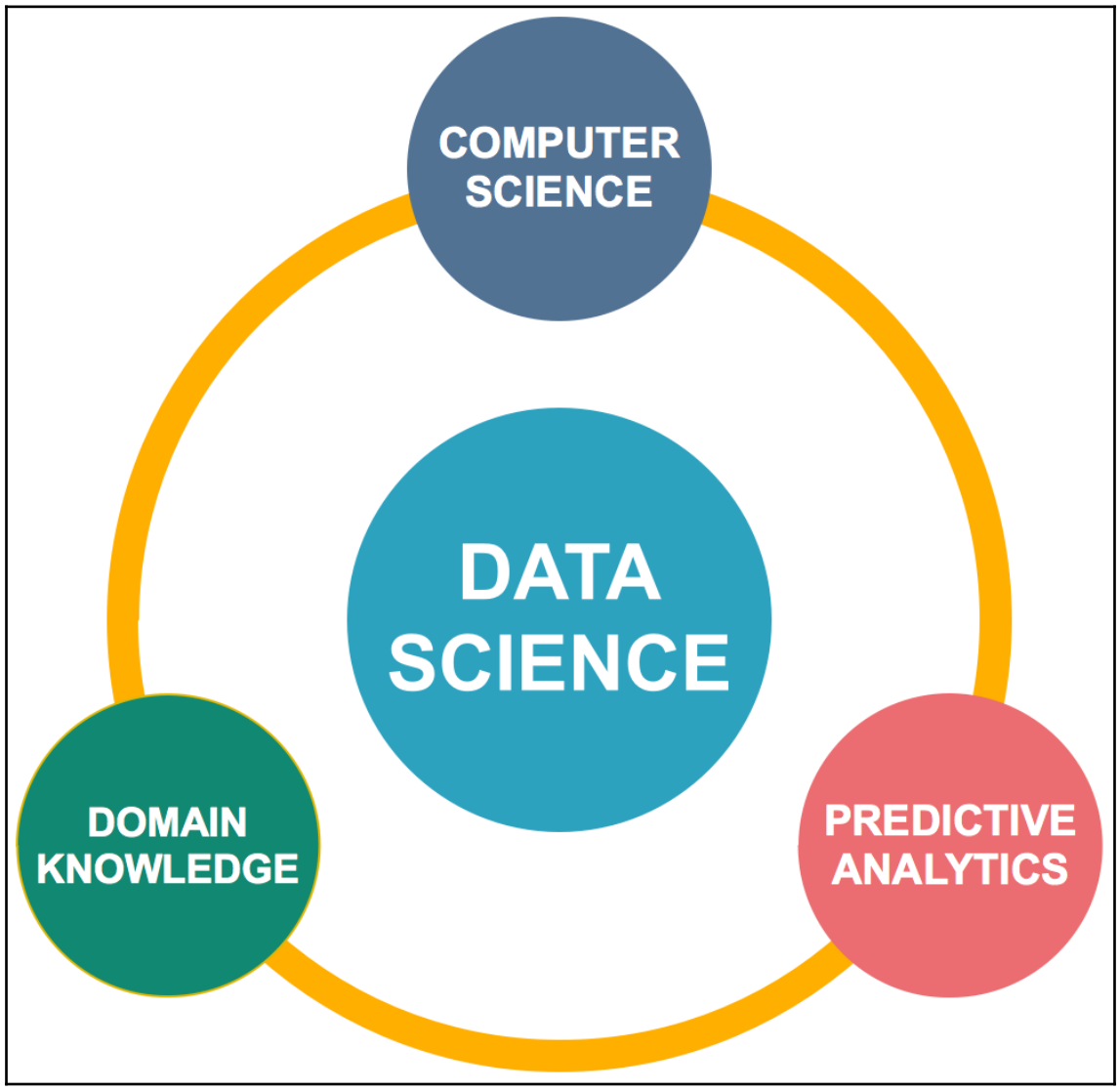
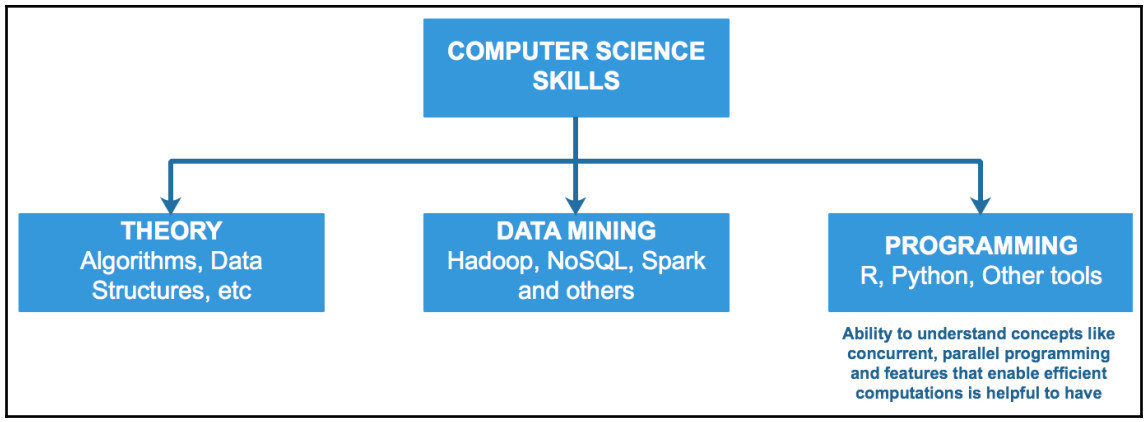
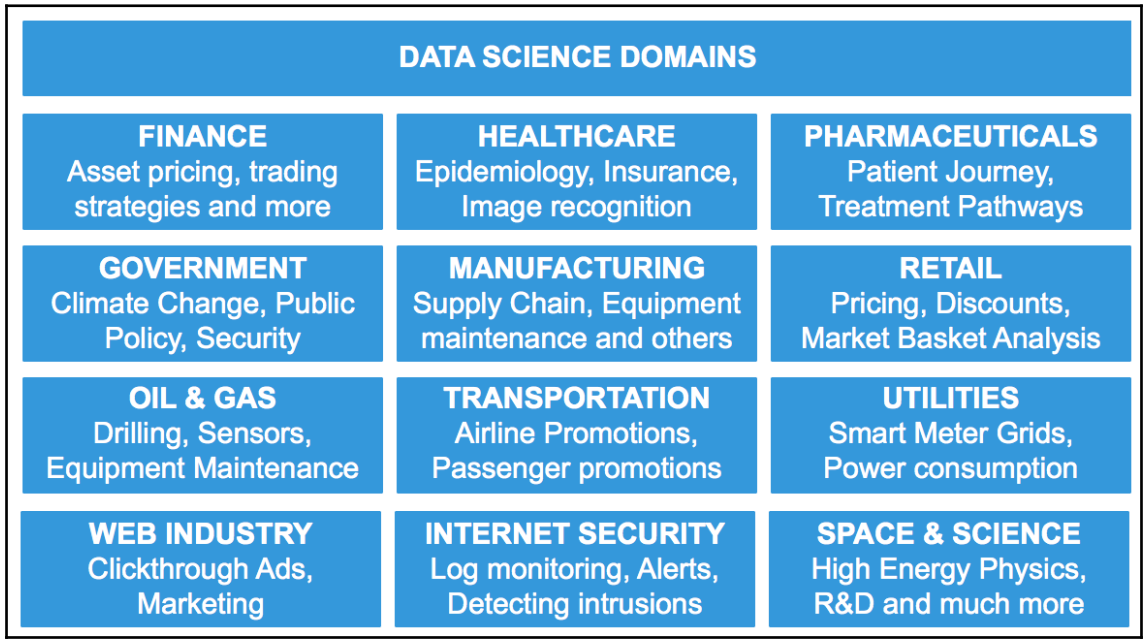
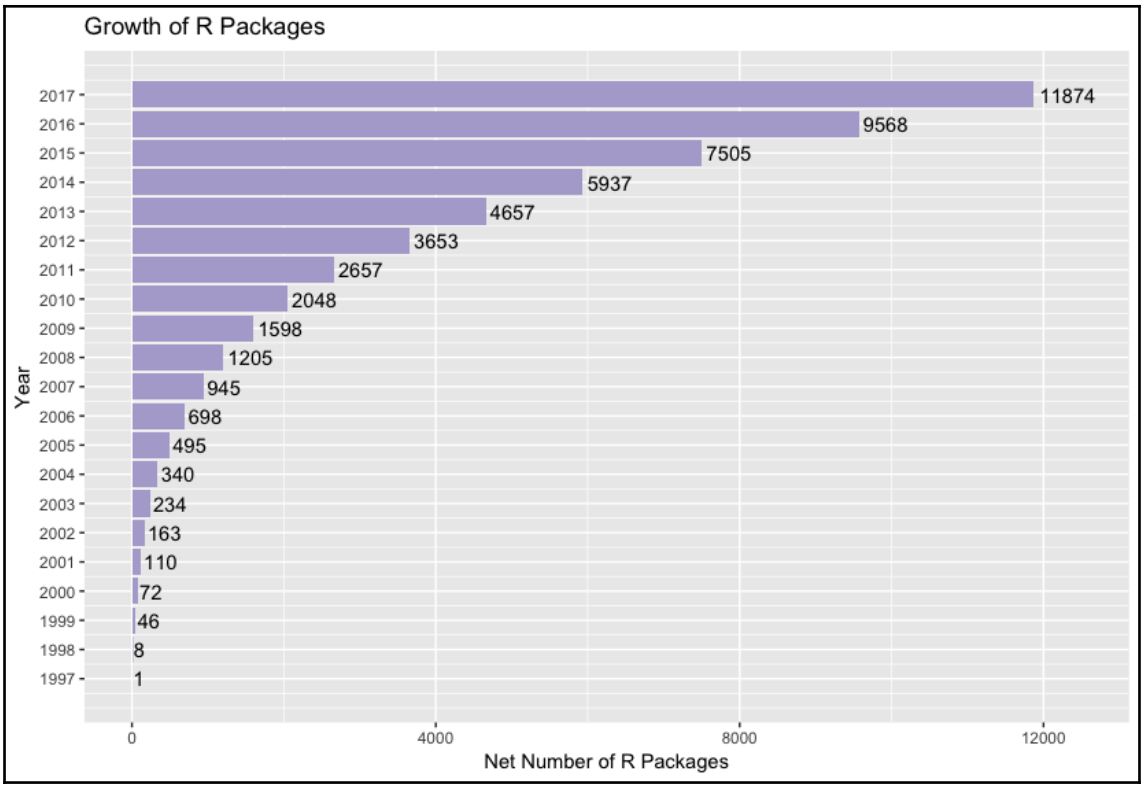
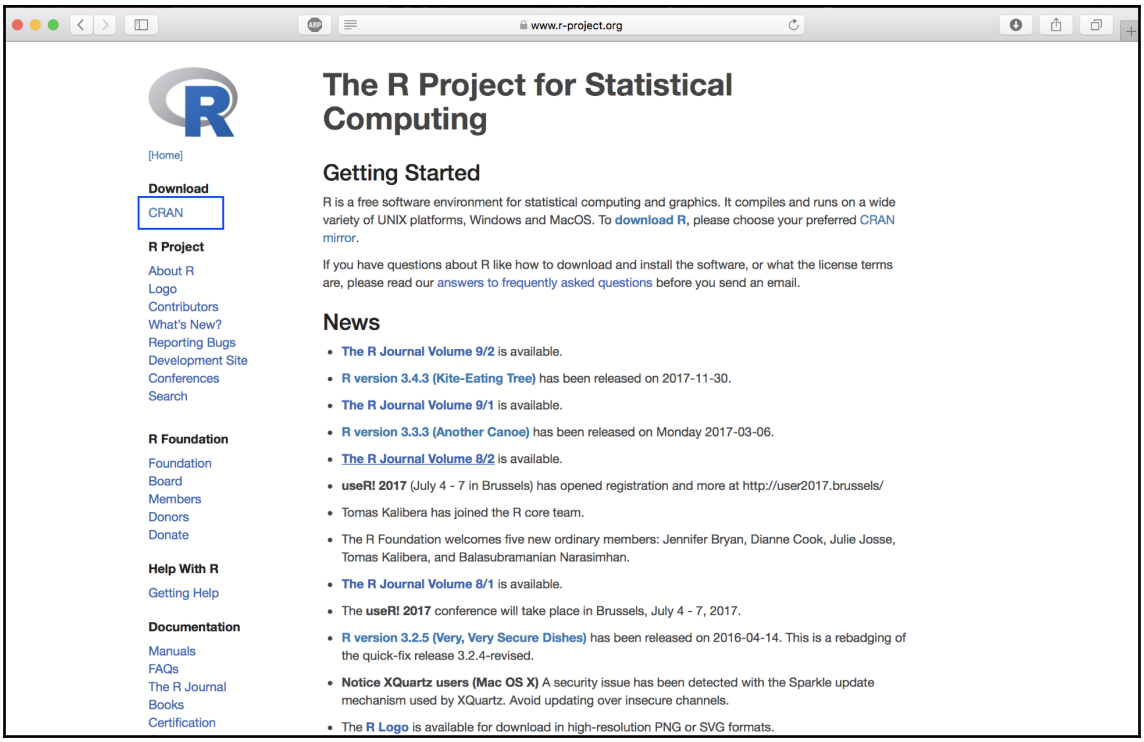
Chapter 1: Getting Started with Data Science and R

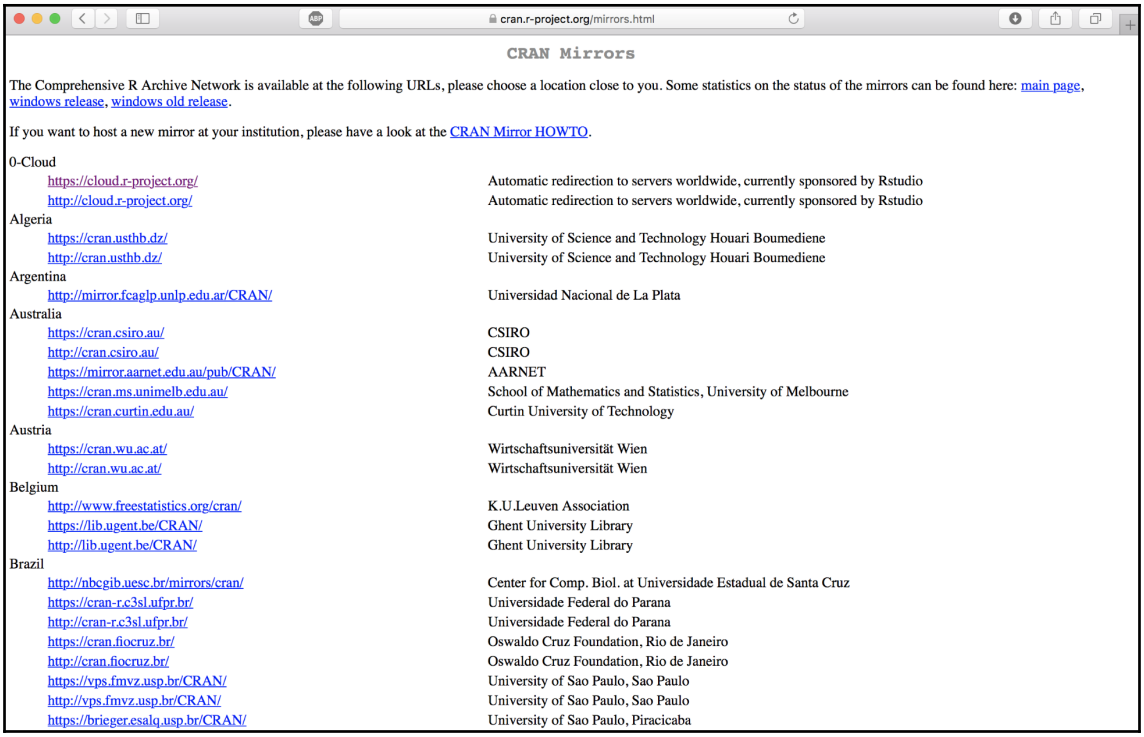


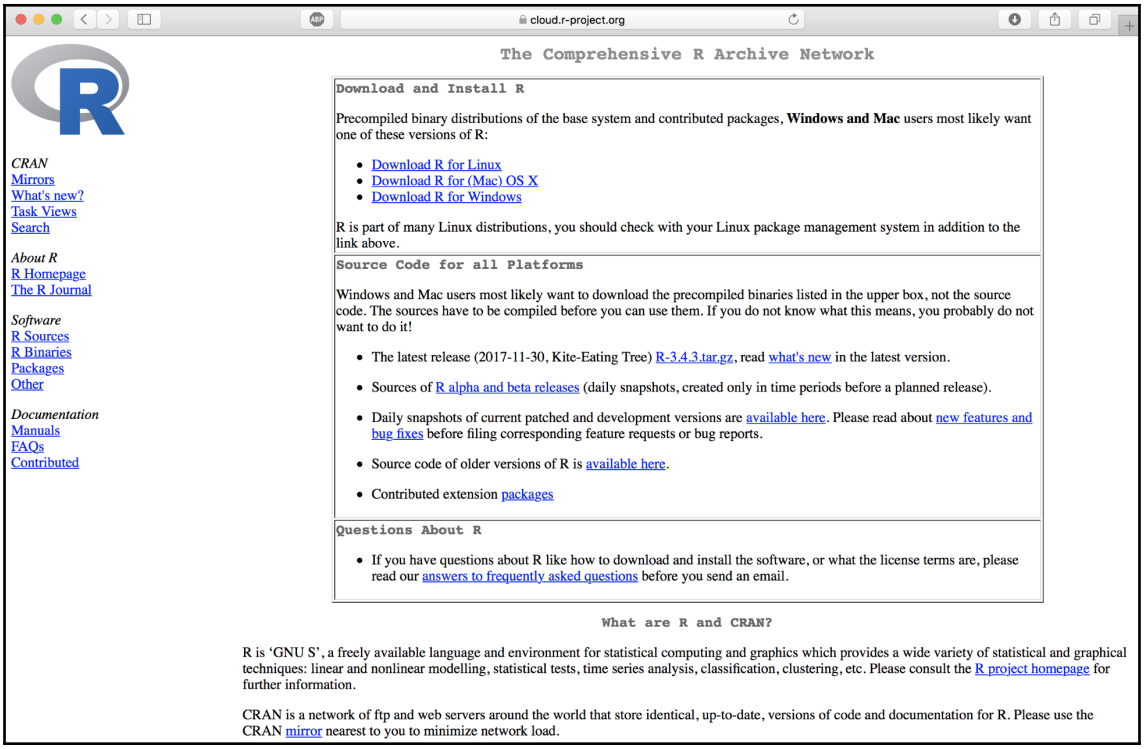

COMPUTER SCIENCE SKILLS
THEORY
Algorithms, Data Structures, etc
DATA MINING
Hadoop, NoSQL, Spark and others
PROGRAMMING
R, Python, Other tools
Ability to understand concepts like concurrent, parallel programming and features that enable efficient computations is helpful to have

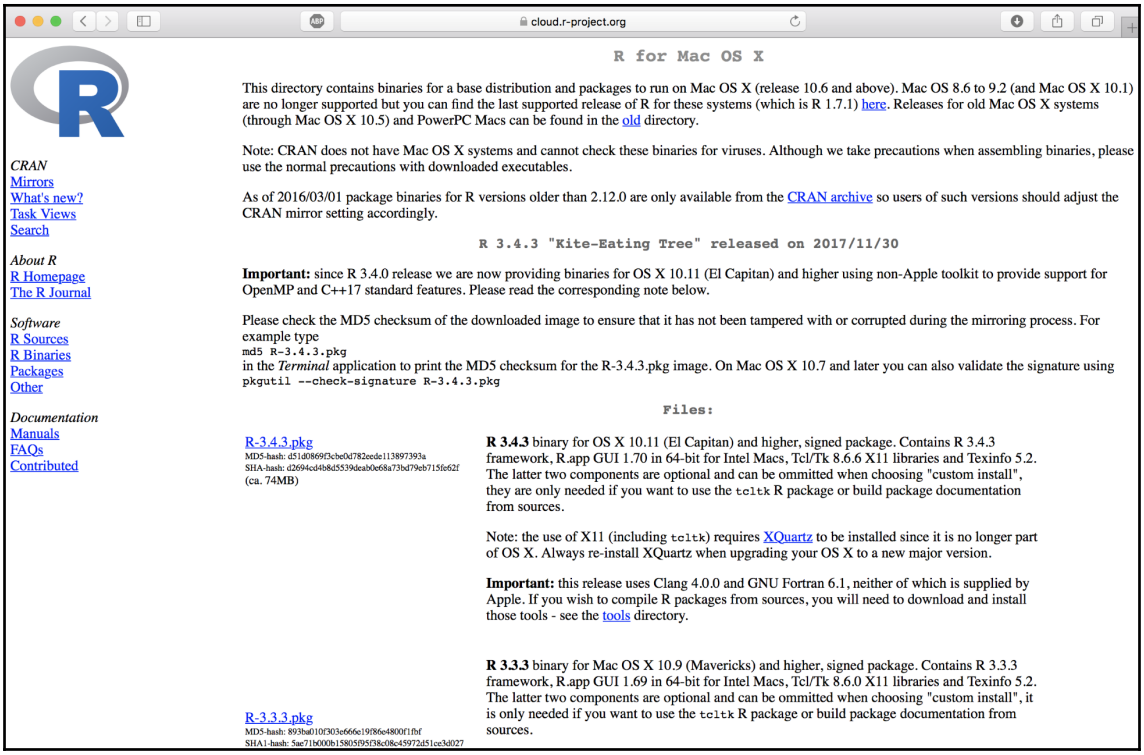

DATA SCIENCE DOMAINS
FINANCE
Asset pricing, trading strategies and more
HEALTHCARE
Epidemiology, Insurance, Image recognition
PHARMACEUTICALS
Patient Journey, Treatment Pathways
GOVERNMENT
Climate Change, Public Policy, Security
MANUFACTURING
Supply Chain, Equipment maintenance and others
RETAIL
Pricing, Discounts, Market Basket Analysis
OIL & GAS
Drilling, Sensors, Equipment Maintenance
TRANSPORTATION
Airline Promotions, Passenger promotions
UTILITIES
Smart Meter Grids, Power consumption
WEB INDUSTRY
Clickthrough Ads, Marketing
INTERNET SECURITY
Log monitoring, Alerts, Detecting intrusions
SPACE & SCIENCE
High Energy Physics, R&D and much more


Growth of R Packages
Year
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999
1998
1997
11874
9568
7505
5937
4657
3653
2657
2048
1598
1205
945
698
495
340
234
163
110
72
46
8
1
0
4000
8000
12000
Net Number of R Packages


www.r-project.org
The R Project for Statistical Computing
[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
Documentation
Manuals
FAQs
The R Journal
Books
Certification
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
The R Journal Volume 9/2 is available.
R version 3.4.3 (Kite-Eating Tree) has been released on 2017-11-30.
The R Journal Volume 9/1 is available.
R version 3.3.3 (Another Canoe) has been released on Monday 2017-03-06.
The R Journal Volume 8/2 is available.
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.
The R Logo is available for download in high-resolution PNG or SVG formats.


cran.r-project.org/mirrors.html
CRAN Mirrors
The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.
If you want to host a new mirror at your institution, please have a look at the CRAN Mirror HOWTO.
0-Cloud
https://cloud.r-project.org/ Automatic redirection to servers worldwide, currently sponsored by Rstudio
http://cloud.r-project.org/ Automatic redirection to servers worldwide, currently sponsored by Rstudio
Algeria
https://cran.usthb.dz/ University of Science and Technology Houari Boumediene
http://cran.usthb.dz/ University of Science and Technology Houari Boumediene
Argentina
http://mirror.fcaglp.unlp.edu.ar/CRAN/ Universidad Nacional de La Plata
Australia
https://cran.csiro.au/ CSIRO
http://cran.csiro.au/ CSIRO
https://mirror.aarnet.edu.au/pub/CRAN/ AARNET
https://cran.ms.unimelb.edu.au/ School of Mathematics and Statistics, University of Melbourne
https://cran.curtin.edu.au/ Curtin University of Technology
Austria
https://cran.wu.ac.at/ Wirtschaftsuniversität Wien
http://cran.wu.ac.at/ Wirtschaftsuniversität Wien
Belgium
http://www.freestatistics.org/cran/ K.U.Leuven Association
https://lib.ugent.be/CRAN/ Ghent University Library
http://lib.ugent.be/CRAN/ Ghent University Library
Brazil
http://nbcgib.uesc.br/mirrors/cran/ Center for Comp. Biol. at Universidade Estadual de Santa Cruz
https://cran-r.c3sl.ufpr.br/ Universidade Federal do Parana
http://cran-r.c3sl.ufpr.br/ Universidade Federal do Parana
https://cran.fiocruz.br/ Oswaldo Cruz Foundation, Rio de Janeiro
http://cran.fiocruz.br/ Oswaldo Cruz Foundation, Rio de Janeiro
https://vps.fmvz.usp.br/CRAN/ University of Sao Paulo, Sao Paulo
http://vps.fmvz.usp.br/CRAN/ University of Sao Paulo, Sao Paulo
https://brieger.esalq.usp.br/CRAN/ University of Sao Paulo, Piracicaba


cloud.r-project.org
The Comprehensive R Archive Network
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
Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Download R for Linux
Download R for (Mac) OS X
Download R for Windows
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.
Source Code for all Platforms
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
The latest release (2017-11-30, Kite-Eating Tree) R-3.4.3.tar.gz, read what's new in the latest version.
Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
Source code of older versions of R is available here.
Contributed extension packages
Questions About R
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.
What are R and CRAN?
R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.
CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load.



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

R for Mac OS X

This directory contains binaries for a base distribution and packages to run on Mac OS X (release 10.6 and above). Mac OS 8.6 to 9.2 (and Mac OS X 10.1) are no longer supported but you can find the last supported release of R for these systems (which is R 1.7.1) here. Releases for old Mac OS X systems (through Mac OS X 10.5) and PowerPC Macs can be found in the old directory.

Note: CRAN does not have Mac OS X systems and cannot check these binaries for viruses. Although we take precautions when assembling binaries, please use the normal precautions with downloaded executables.

As of 2016/03/01 package binaries for R versions older than 2.12.0 are only available from the CRAN archive so users of such versions should adjust the CRAN mirror setting accordingly.

R 3.4.3 "Kite-Eating Tree" released on 2017/11/30

Important: since R 3.4.0 release we are now providing binaries for OS X 10.11 (El Capitan) and higher using non-Apple toolkit to provide support for OpenMP and C++17 standard features. Please read the corresponding note below.

Please check the MD5 checksum of the downloaded image to ensure that it has not been tampered with or corrupted during the mirroring process. For example type
`md5 R-3.4.3.pkg`
in the *Terminal* application to print the MD5 checksum for the R-3.4.3.pkg image. On Mac OS X 10.7 and later you can also validate the signature using `pkgutil --check-signature R-3.4.3.pkg`

Files:

R-3.4.3.pkg
MD5-hash: d51d0869f3cbe0d782eede113897393a
SHA-hash: d2694cd4b8d5539deab0e68a73bd79eb715fe62f
(ca. 74MB)

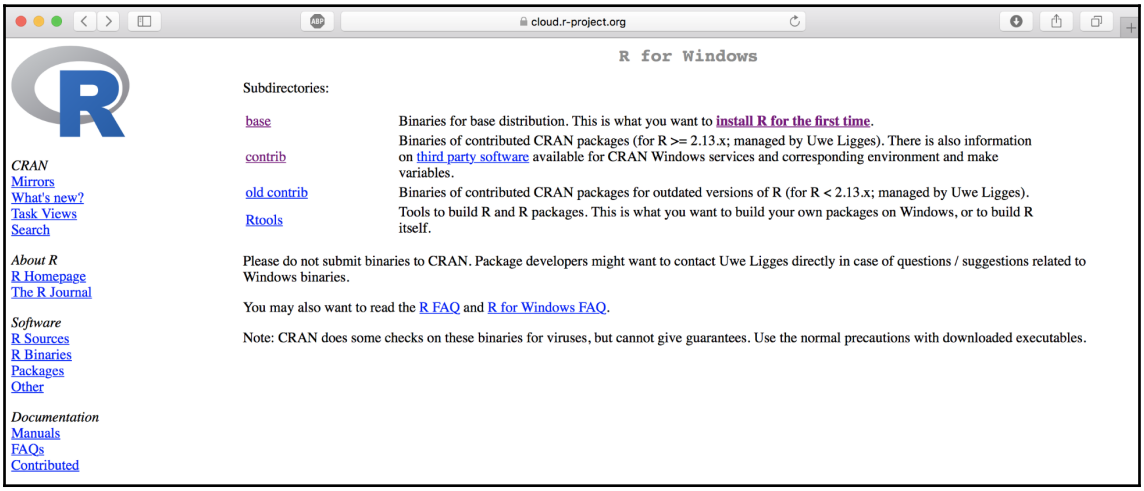
R 3.4.3 binary for OS X 10.11 (El Capitan) and higher, signed package. Contains R 3.4.3 framework, R.app GUI 1.70 in 64-bit for Intel Macs, Tcl/Tk 8.6.6 X11 libraries and Texinfo 5.2. The latter two components are optional and can be ommitted when choosing "custom install", they are only needed if you want to use the `tcltk` R package or build package documentation from sources.

Note: the use of X11 (including `tcltk`) requires XQuartz to be installed since it is no longer part of OS X. Always re-install XQuartz when upgrading your OS X to a new major version.

Important: this release uses Clang 4.0.0 and GNU Fortran 6.1, neither of which is supplied by Apple. If you wish to compile R packages from sources, you will need to download and install those tools - see the tools directory.

R-3.3.3.pkg
MD5-hash: 893ba010f303e666e19f86e4800f1fbf
SHA1-hash: 5ae71b000b15805f95f38c08c45972d51ce3d027

R 3.3.3 binary for Mac OS X 10.9 (Mavericks) and higher, signed package. Contains R 3.3.3 framework, R.app GUI 1.69 in 64-bit for Intel Macs, Tcl/Tk 8.6.0 X11 libraries and Texinfo 5.2. The latter two components are optional and can be ommitted when choosing "custom install", it is only needed if you want to use the `tcltk` R package or build package documentation from sources.



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

R for Windows

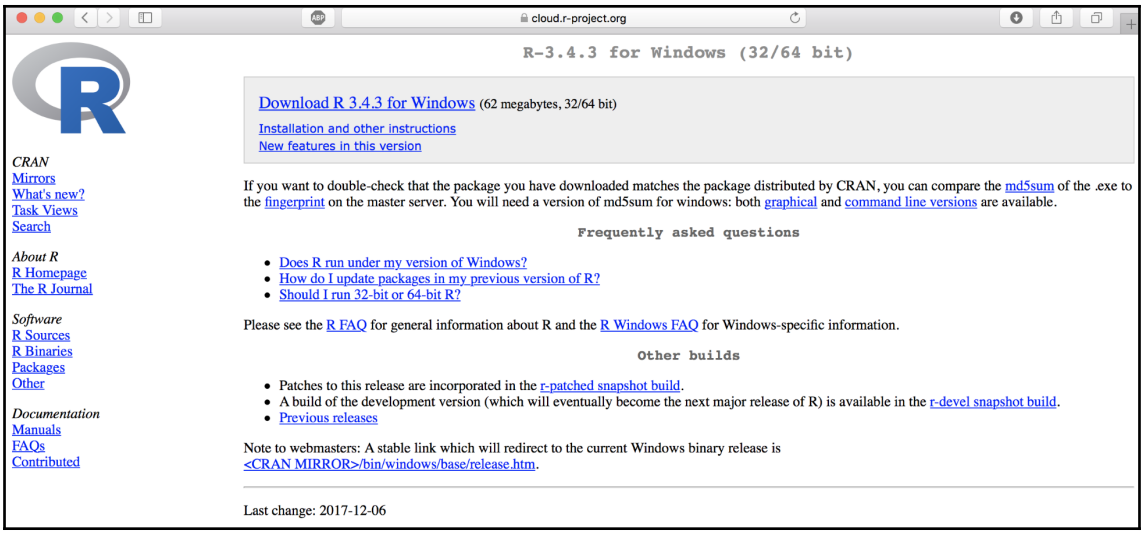
Subdirectories:

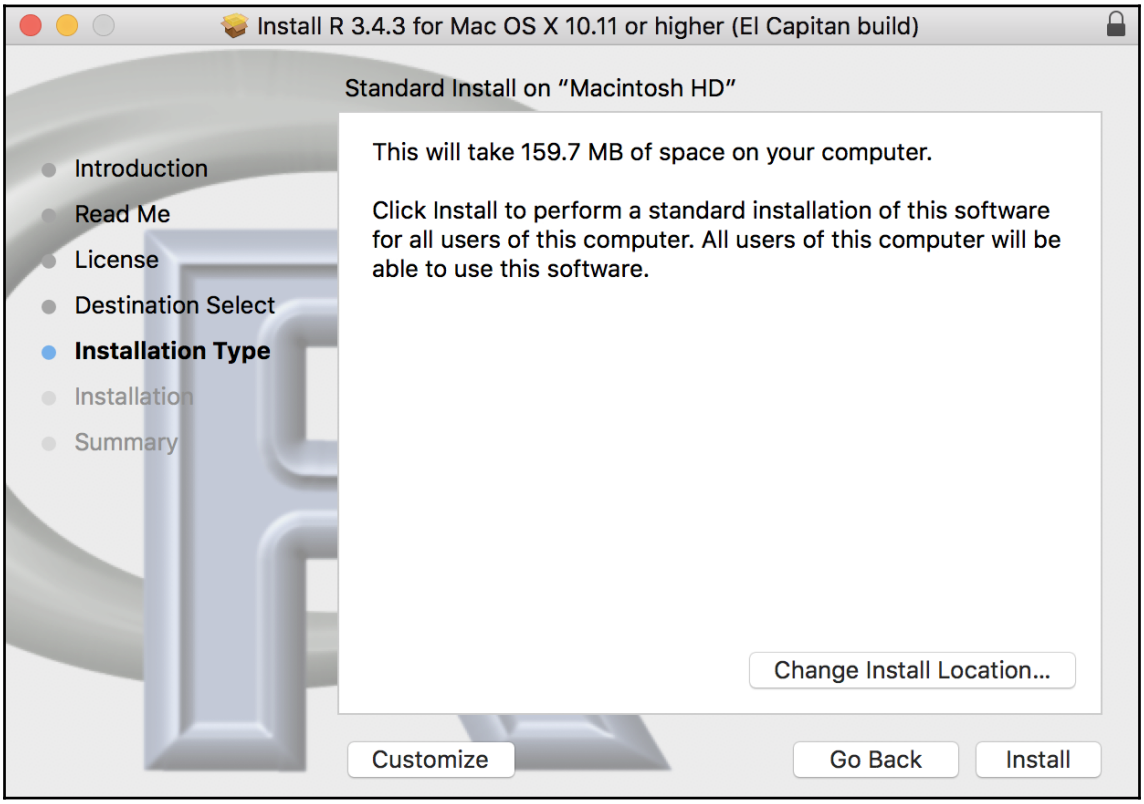
base	Binaries for base distribution. This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

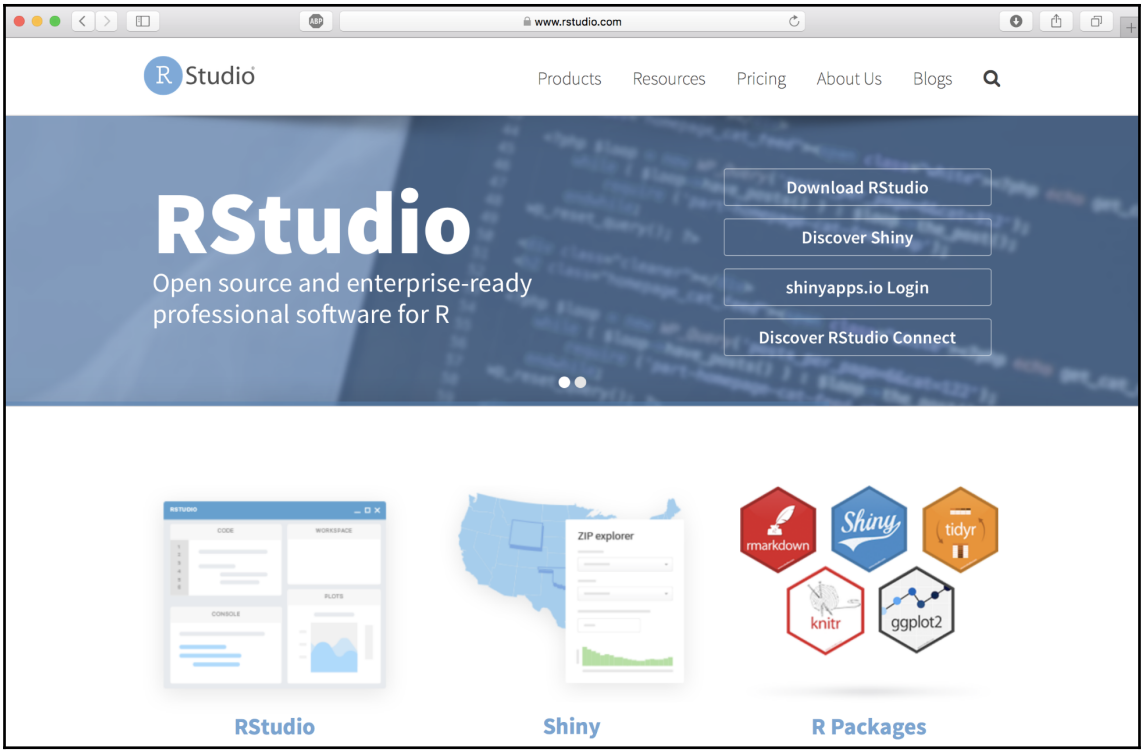
Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

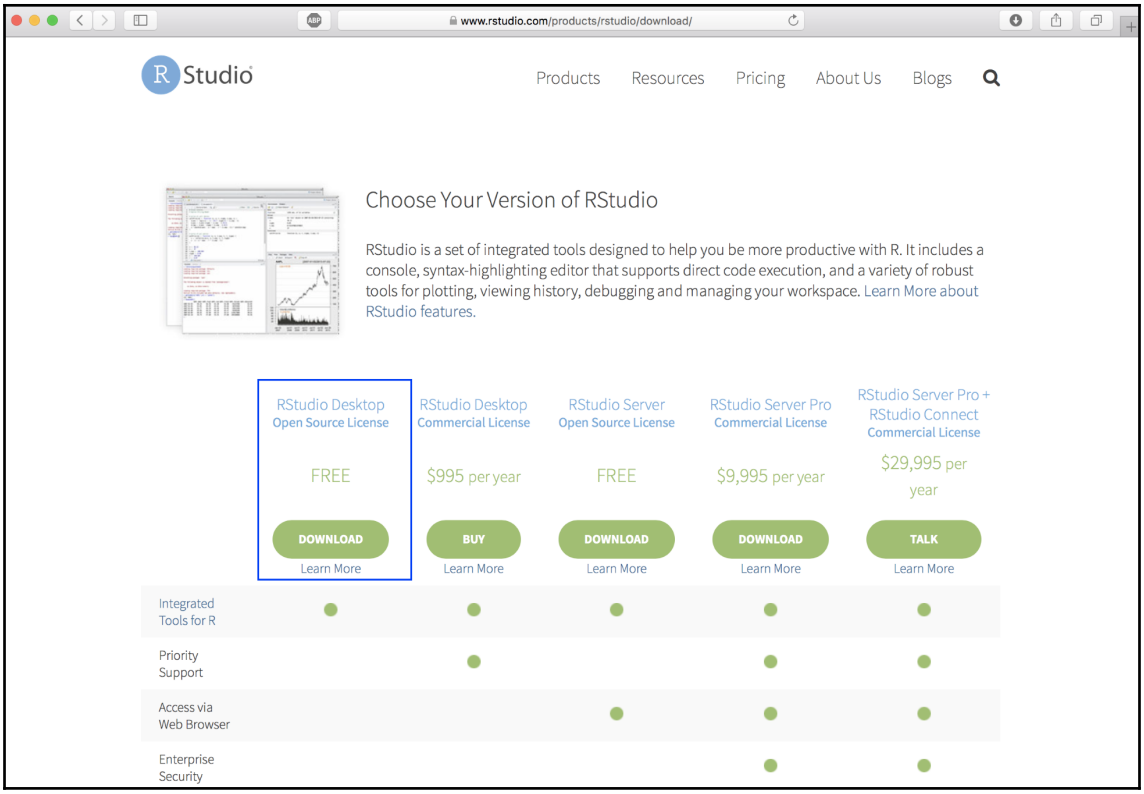
You may also want to read the R FAQ and R for Windows FAQ.

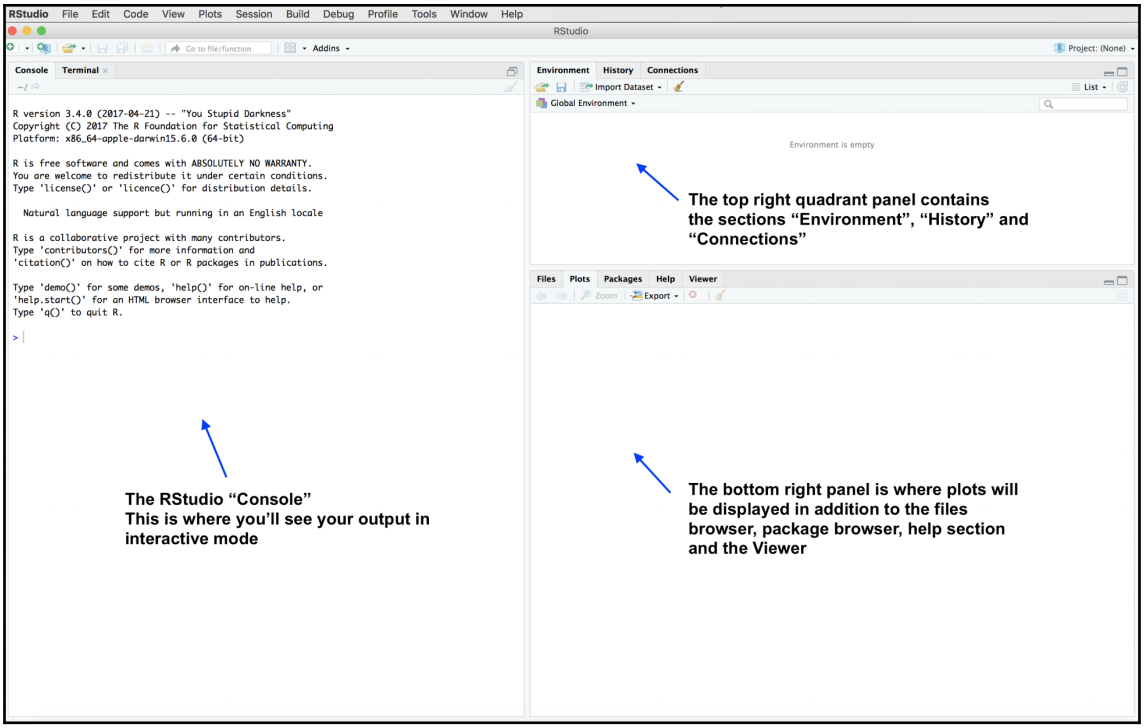
Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

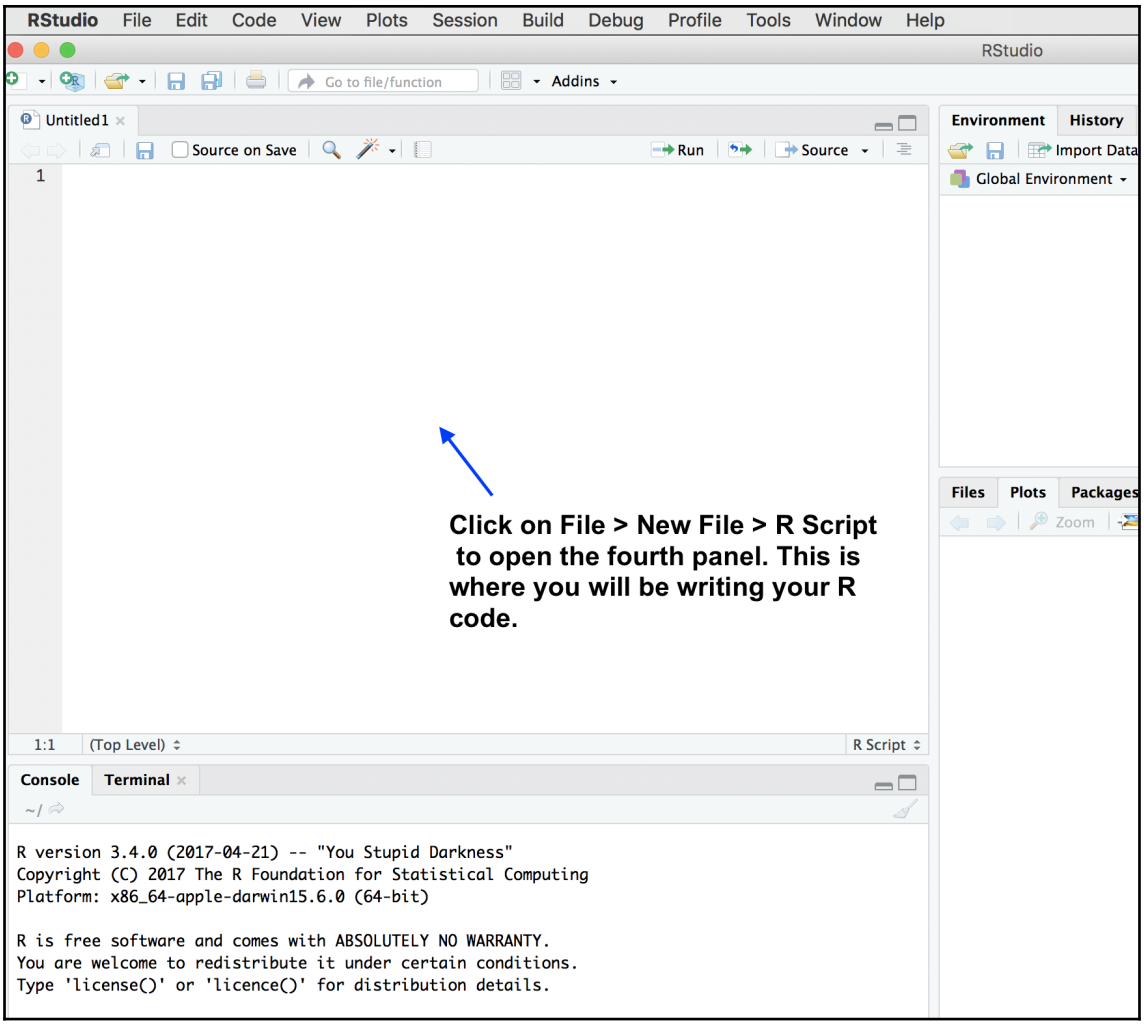

cloud.r-project.org
R-3.4.3 for Windows (32/64 bit)
Download R 3.4.3 for Windows (62 megabytes, 32/64 bit)
Installation and other instructions
New features in this version
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.
Frequently asked questions
Does R run under my version of Windows?
How do I update packages in my previous version of R?
Should I run 32-bit or 64-bit R?
Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.
Other builds
Patches to this release are incorporated in the r-patched snapshot build.
A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
Previous releases
Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.
Last change: 2017-12-06

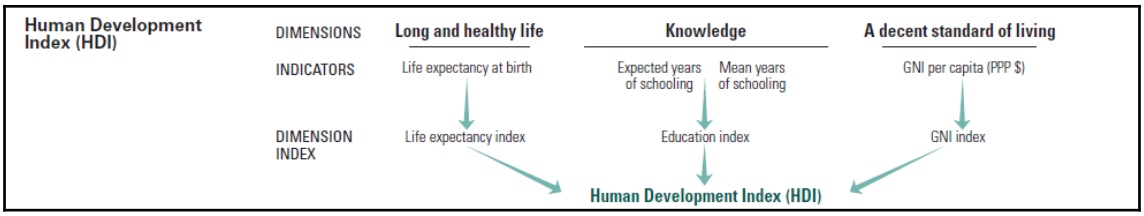

Install R 3.4.3 for Mac OS X 10.11 or higher (El Capitan build)
Standard Install on "Macintosh HD"
Introduction
Read Me
License
Destination Select
Installation Type
Installation
Summary
This will take 159.7 MB of space on your computer.
Click Install to perform a standard installation of this software for all users of this computer. All users of this computer will be able to use this software.
Change Install Location...
Customize
Go Back
Install

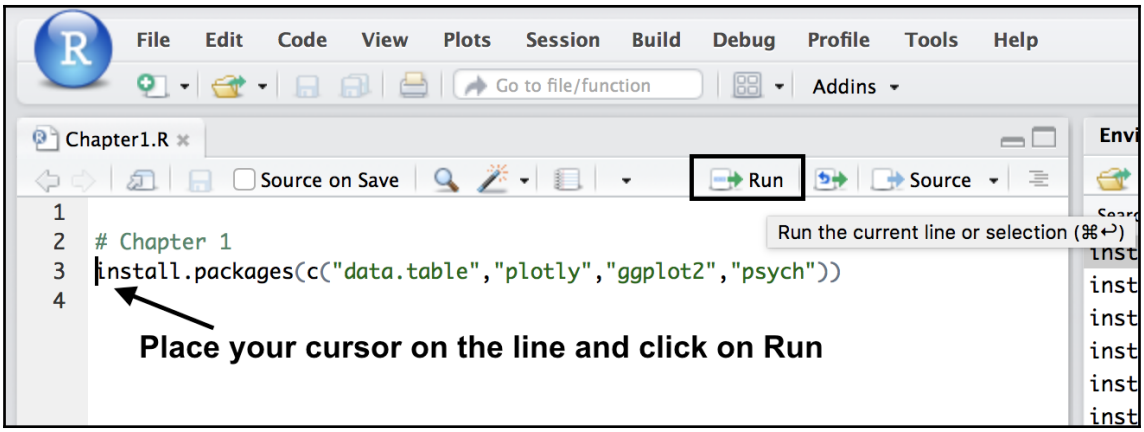

www.rstudio.com
RStudio
Products
Resources
Pricing
About Us
Blogs
RStudio
Open source and enterprise-ready professional software for R
Download RStudio
Discover Shiny
shinyapps.io Login
Discover RStudio Connect
ZIP explorer
rmarkdown
Shiny
tidyr
knitr
ggplot2
RStudio
Shiny
R Packages


www.rstudio.com/products/rstudio/download/
RStudio
Products
Resources
Pricing
About Us
Blogs
Choose Your Version of RStudio
RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. Learn More about RStudio features.
RStudio Desktop
Open Source License
FREE
DOWNLOAD
Learn More
RStudio Desktop
Commercial License
$995 per year
BUY
Learn More
RStudio Server
Open Source License
FREE
DOWNLOAD
Learn More
RStudio Server Pro
Commercial License
$9,995 per year
DOWNLOAD
Learn More
RStudio Server Pro + RStudio Connect
Commercial License
$29,995 per year
TALK
Learn More
Integrated Tools for R
Priority Support
Access via Web Browser
Enterprise Security


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
RStudio
Console Terminal
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
Environment History Connections
Import Dataset
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Export
The top right quadrant panel contains the sections “Environment”, “History” and “Connections”
The RStudio “Console” This is where you’ll see your output in interactive mode
The bottom right panel is where plots will be displayed in addition to the files browser, package browser, help section and the Viewer


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
RStudio
Go to file/function
Addins
Untitled1
Source on Save
Run
Source
Environment
History
Import Data
Global Environment
Files
Plots
Packages
Zoom
Click on File > New File > R Script to open the fourth panel. This is where you will be writing your R code.
1:1
(Top Level)
R Script
Console
Terminal
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.


Human Development Index (HDI)
DIMENSIONS
INDICATORS
DIMENSION INDEX
Long and healthy life
Life expectancy at birth
Life expectancy index
Knowledge
Expected years of schooling
Mean years of schooling
Education index
A decent standard of living
GNI per capita (PPP $)
GNI index
Human Development Index (HDI)




```
> head(hdi) # View the top few rows of the data table hdi
                Country Year   HDI
1:          Afghanistan 1990 0.295
2:              Albania 1990 0.635
3:              Algeria 1990 0.577
4:              Andorra 1990    NA
5:               Angola 1990    NA
6: Antigua and Barbuda 1990    NA
```

```
> head(life)
                 Country Year LifeExp
1:           Afghanistan 1990    49.9
2:               Albania 1990    71.8
3:               Algeria 1990    66.7
4:               Andorra 1990    76.5
5:                Angola 1990    41.2
6: Antigua and Barbuda 1990    71.4
```

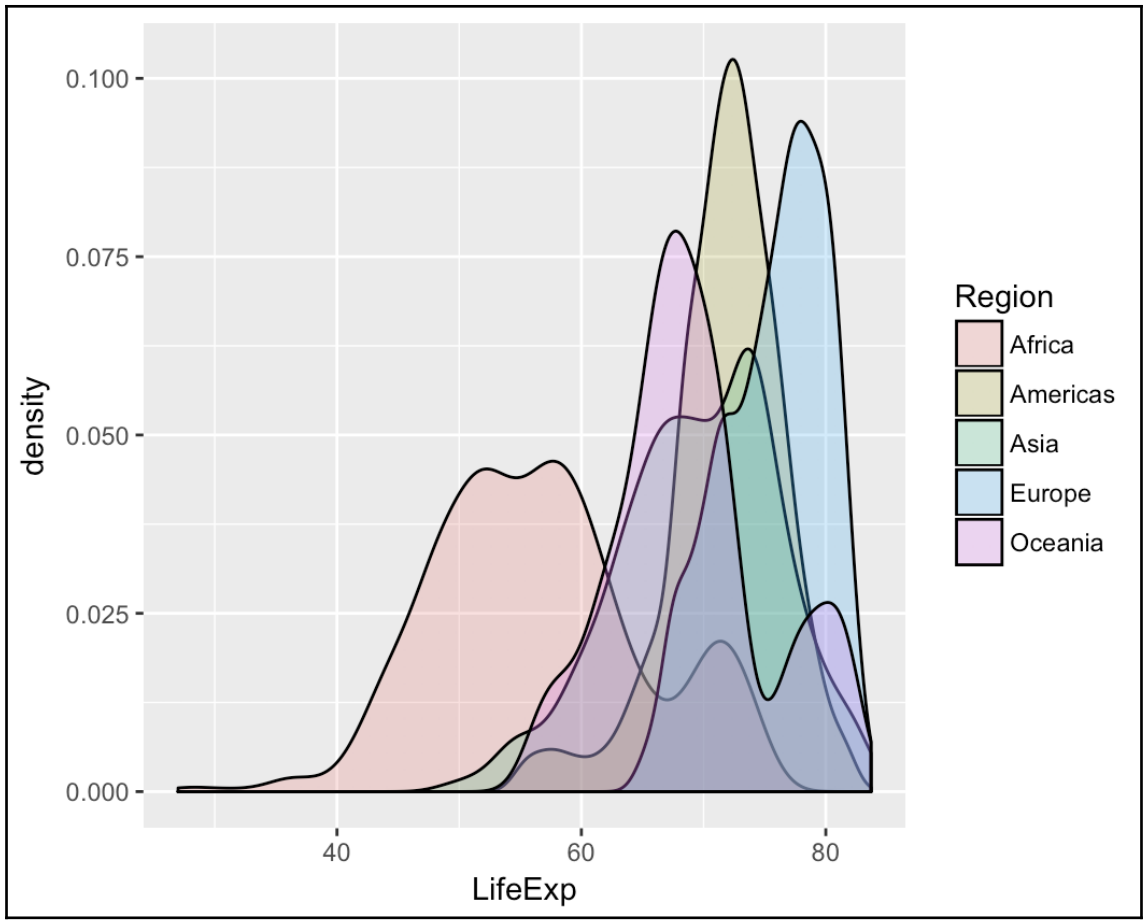
```
> head(school)
                 Country Year SchoolYrs
1:           Afghanistan 1990       2.6
2:               Albania 1990      11.6
3:               Algeria 1990       9.6
4:               Andorra 1990      10.8
5:                Angola 1990       3.8
6: Antigua and Barbuda 1990        NA
```

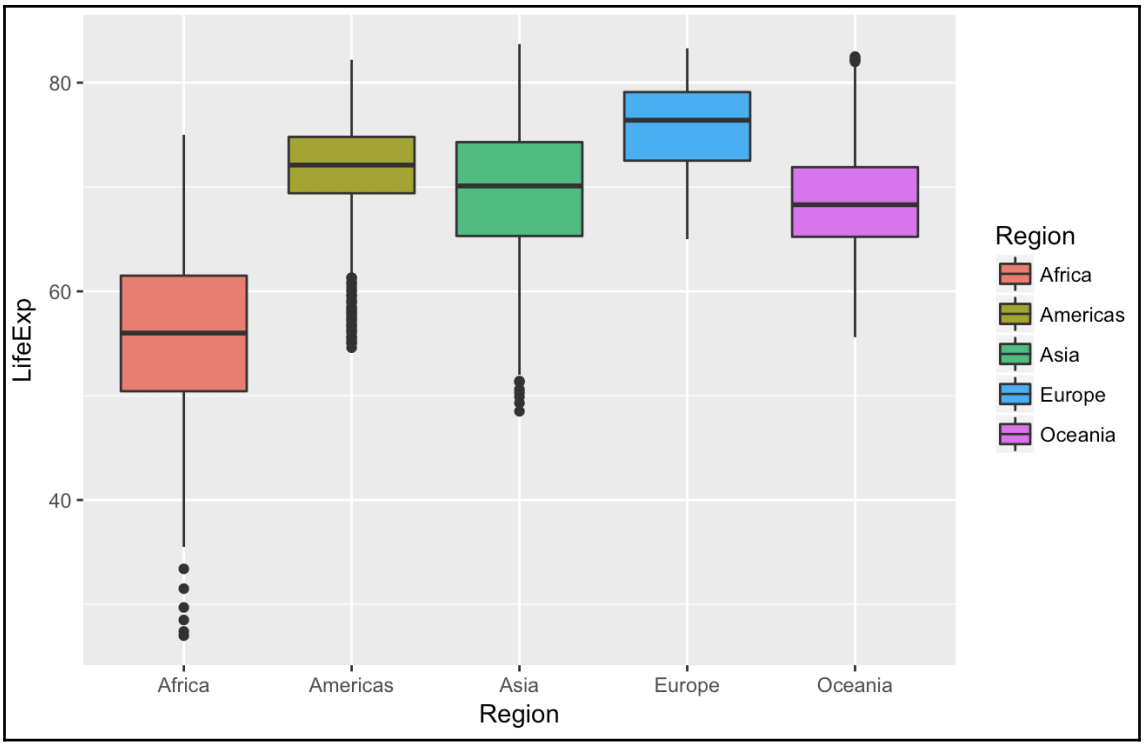
```
> head(iso)
          Country  Region       SubRegion
1:    Afghanistan    Asia   Southern Asia
2:  Åland Islands  Europe Northern Europe
3:        Albania  Europe Southern Europe
4:        Algeria  Africa Northern Africa
5: American Samoa Oceania       Polynesia
6:        Andorra  Europe Southern Europe
```

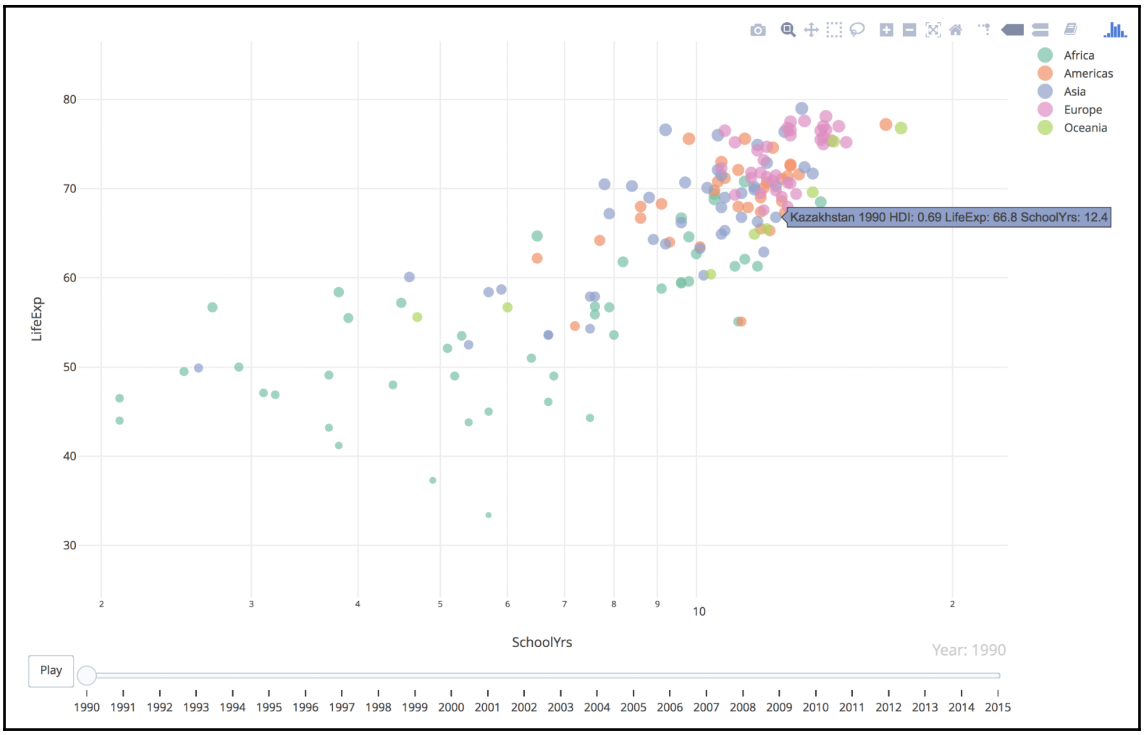
Filter

	Country	Year	HDI	LifeExp	SchoolYrs	Region	SubRegion	Info
1	Afghanistan	1990	0.295	49.9	2.6	Asia	Southern Asia	Afghanistan 1990 HDI: 0.295 LifeExp: 49.9 SchoolYrs...
2	Afghanistan	1991	0.300	50.6	2.9	Asia	Southern Asia	Afghanistan 1991 HDI: 0.3 LifeExp: 50.6 SchoolYrs: 2.9
3	Afghanistan	1992	0.309	51.4	3.2	Asia	Southern Asia	Afghanistan 1992 HDI: 0.309 LifeExp: 51.4 SchoolYrs...
4	Afghanistan	1993	0.305	52.0	3.6	Asia	Southern Asia	Afghanistan 1993 HDI: 0.305 LifeExp: 52 SchoolYrs: 3.6
5	Afghanistan	1994	0.300	52.6	3.9	Asia	Southern Asia	Afghanistan 1994 HDI: 0.3 LifeExp: 52.6 SchoolYrs: 3.9
6	Afghanistan	1995	0.324	53.1	4.2	Asia	Southern Asia	Afghanistan 1995 HDI: 0.324 LifeExp: 53.1 SchoolYrs...

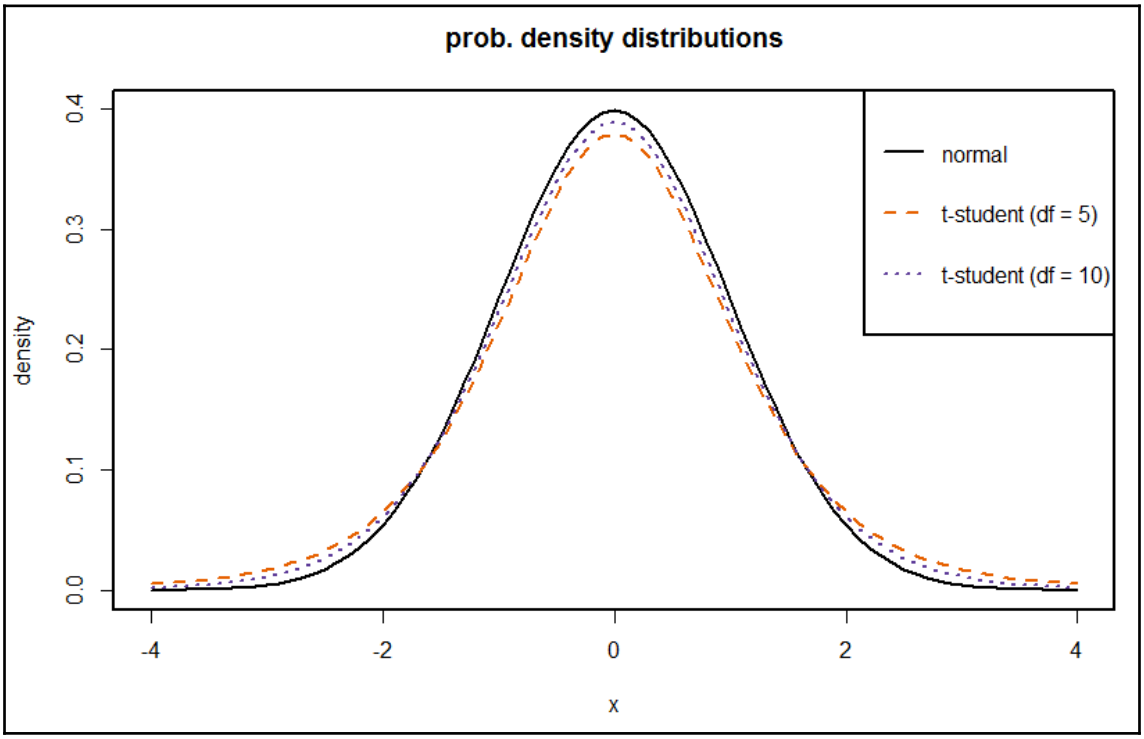
```
> mergedDataSummary["Cuba"] # Enter any country name here to view the summary information
$Cuba
           vars  n  mean   sd median trimmed  mad   min   max range  skew kurtosis   se IQR
HDI           1 26  0.72 0.05   0.70    0.72 0.06  0.65  0.78  0.13  0.17    -1.77 0.01 0.1
LifeExp       2 26 77.21 1.68  77.25   77.24 2.30 74.60 79.60  5.00 -0.12    -1.52 0.33 3.0
SchoolYrs     3 26 13.69 2.00  13.05   13.54 1.70 11.30 17.70  6.40  0.68    -0.91 0.39 2.6
```


0.100
0.075
0.050
0.025
0.000
density
40
60
80
LifeExp
Region
Africa
Americas
Asia
Europe
Oceania

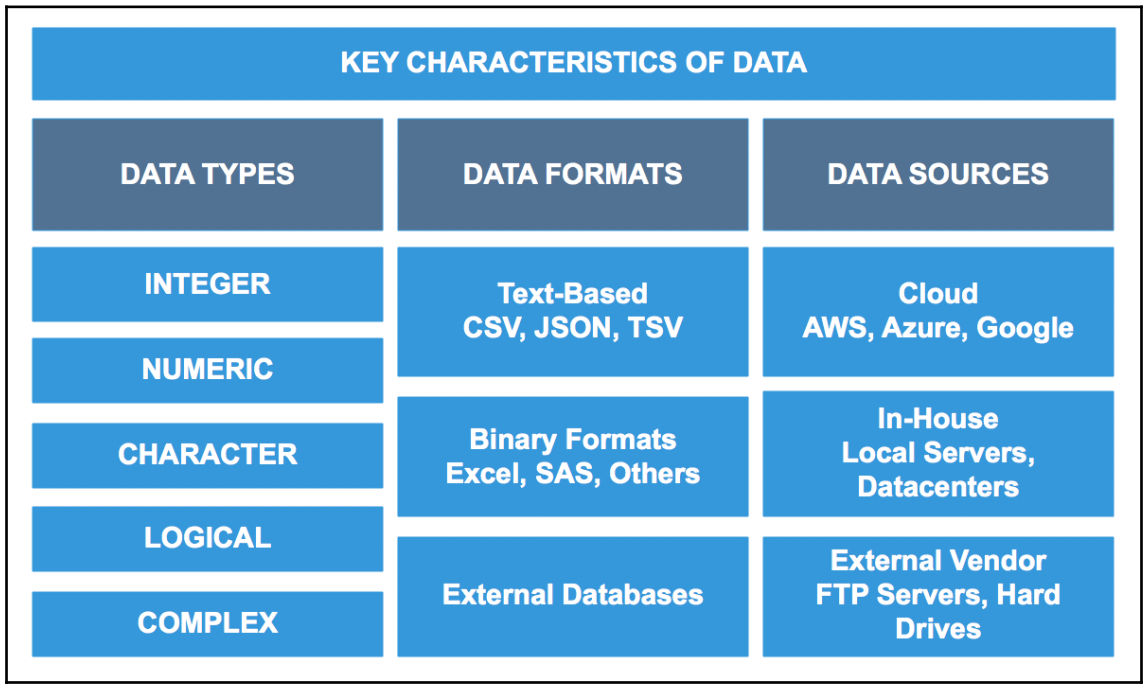

80
60
40
LifeExp
Africa
Americas
Asia
Europe
Oceania
Region
Region
Africa
Americas
Asia
Europe
Oceania


Africa
Americas
Asia
Europe
Oceania
Kazakhstan 1990 HDI: 0.69 LifeExp: 66.8 SchoolYrs: 12.4
LifeExp
SchoolYrs
Year: 1990
Play
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

Chapter 2: Descriptive and Inferential Statistics



Chapter 3: Data Wrangling with R



Data Frame

COLUMN NAMES (colnames or names)

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56.0	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2.0	68.54	13.9	40.6	60	58073
Hawaii	868	4963	1.9	73.60	6.2	61.9	0	6425
Idaho	813	4119	0.6	71.87	5.3	59.5	126	82677
Illinois	11197	5107	0.9	70.14	10.3	52.6	127	55748
Indiana	5313	4458	0.7	70.88	7.1	52.9	122	36097

ROWNAMES (row.names())

```
> str(state)
'data.frame':   50 obs. of  9 variables:
 $ Population: num  3615 365 2212 2110 21198 ...
 $ Income    : num  3624 6315 4530 3378 5114 ...
 $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
 $ Life.Exp  : num  69 69.3 70.5 70.7 71.7 ...
 $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
 $ HS.Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
 $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
 $ Area      : num  50708 566432 113417 51945 156361 ...
 $ State     : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
```

```
> apply(state[,-ncol(state)], 2, mean) # Mean of all values in the numeric columns
Population     Income Illiteracy   Life.Exp     Murder    HS.Grad      Frost       Area
 4246.4200  4435.8000     1.1700    70.8786     7.3780    53.1080  104.4600 70735.8800
```

```
> lapply(state[,-ncol(state)], function(x) {list(MIN=min(x), MAX=max(x), MEAN=mean(x))})
$Population
$Population$MIN
[1] 365

$Population$MAX
[1] 21198

$Population$MEAN
[1] 4246.42


$Income
$Income$MIN
[1] 3098
```

```
> sapply(state[,-ncol(state)], function(x) {list(MIN=min(x), MAX=max(x), MEAN=mean(x))})
     Population Income Illiteracy Life.Exp Murder HS.Grad Frost  Area
MIN  365        3098   0.5        67.96    1.4    37.8    0      1049
MAX  21198      6315   2.8        73.6     15.1   67.3    188    566432
MEAN 4246.42    4435.8 1.17       70.8786  7.378  53.108  104.46 70735.88
```

```
> aggregate(state[,-c(9,10)], by=list(state$Region), mean, na.rm = T)
        Group.1 Population   Income Illiteracy Life.Exp    Murder  HS.Grad    Frost      Area
1     Northeast   5495.111 4570.222   1.000000 71.26444  4.722222 53.96667 132.7778  18141.00
2         South   4208.125 4011.938   1.737500 69.70625 10.581250 44.34375  64.6250  54605.12
3 North Central   4803.000 4611.083   0.700000 71.76667  5.275000 54.51667 138.8333  62652.00
4          West   2915.308 4702.615   1.023077 71.23462  7.215385 62.00000 102.1538 134463.00
```

Common Key (State)

```
merge(state,state2,by="State",all=T)
```

state data.frame

State	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area	Region
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708	South
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432	West
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417	West
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945	South
California	21198	5114	1.1	71.71	10.3	62.6	20	156361	West
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766	West
Connecticut	3100	5348	1.1	72.48	3.1	56.0	139	4862	Northeast
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982	South

state2 data.frame

Latitude	Longitude
32.5901	-86.7509
49.2500	-127.2500
34.2192	-111.6250
34.7336	-92.2992
36.5341	-119.7730
38.6777	-105.5130
41.5928	-72.3573
38.6777	-74.9841

```
> state01
         Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area    State Region
Alabama        3615   3624        2.1    69.05   15.1    41.3    20  50708  Alabama  South
Alaska          365   6315        1.5    69.31   11.3    66.7   152 566432   Alaska   West
```

```
> library("tidyverse")
── Attaching packages ─────────────────────
✔ ggplot2 2.2.1     ✔ purrr   0.2.4
✔ tibble  1.3.4     ✔ dplyr   0.7.4
✔ tidyr   0.8.0     ✔ stringr 1.3.0
✔ readr   1.1.1     ✔ forcats 0.2.0
```

```
> step1
        Group.1 Population   Income Illiteracy Life Exp    Murder  HS Grad    Frost      Area
1     Northeast   5495.111 4570.222   1.000000 71.26444  4.722222 53.96667 132.7778  18141.00
2         South   4208.125 4011.938   1.737500 69.70625 10.581250 44.34375  64.6250  54605.12
3 North Central   4803.000 4611.083   0.700000 71.76667  5.275000 54.51667 138.8333  62652.00
4          West   2915.308 4702.615   1.023077 71.23462  7.215385 62.00000 102.1538 134463.00
```

```
> step2
  Group.1 Population   Income Illiteracy Life Exp   Murder HS Grad    Frost   Area
4    West   2915.308 4702.615   1.023077 71.23462 7.215385      62 102.1538 134463
```

```
# A tibble: 1 x 9
  Region Population   Income Illiteracy `Life Exp`   Murder `HS Grad`    Frost   Area
  <fctr>      <dbl>    <dbl>      <dbl>      <dbl>    <dbl>      <dbl>    <dbl>  <dbl>
1   West   2915.308 4702.615   1.023077   71.23462 7.215385         62 102.1538 134463
```

```
# A tibble: 4 x 6
         Region total_rows first_state unique_states max_literacy mean_literacy
         <fctr>      <int>       <chr>         <int>        <dbl>          <dbl>
1     Northeast          9 Connecticut             9         99.4       99.00000
2         South         16     Alabama            16         99.1       98.26250
3 North Central         12    Illinois            12         99.5       99.30000
4          West         13      Alaska            13         99.5       98.97692
```

```
> sample_n(tstate, 10) # To select 10 random rows
# A tibble: 10 x 10
   Population Income Illiteracy `Life Exp` Murder `HS Grad` Frost  Area        Region          Name
        <dbl>  <dbl>      <dbl>      <dbl>  <dbl>     <dbl> <dbl> <dbl>        <fctr>         <chr>
 1        637   5087        0.8      72.78    1.4      50.3   186 69273 North Central  North Dakota
 2        868   4963        1.9      73.60    6.2      61.9     0  6425          West        Hawaii
 3       3806   3545        2.8      68.76   13.2      42.2    12 44930         South     Louisiana
 4       3387   3712        1.6      70.10   10.6      38.5    95 39650         South      Kentucky
 5      18076   4903        1.4      70.55   10.9      52.7    82 47831     Northeast      New York
 6       4767   4254        0.8      70.69    9.3      48.8   108 68995 North Central      Missouri
 7       4981   4701        1.4      70.08    9.5      47.8    85 39780         South      Virginia
 8       2341   3098        2.4      68.09   12.5      41.0    50 47296         South   Mississippi
 9       1203   4022        0.6      72.90    4.5      67.3   137 82096          West          Utah
10       1799   3617        1.4      69.48    6.7      41.6   100 24070         South West Virginia
```

```
# A tibble: 50 x 11
   Population Income Illiteracy `Life Exp` Murder `HS Grad` Frost   Area    Region        Name  Abbr
        <dbl>  <dbl>      <dbl>      <dbl>  <dbl>     <dbl> <dbl>  <dbl>    <fctr>       <chr> <chr>
 1       3615   3624        2.1      69.05   15.1      41.3    20  50708     South     Alabama    AL
 2        365   6315        1.5      69.31   11.3      66.7   152 566432      West      Alaska    AK
 3       2212   4530        1.8      70.55    7.8      58.1    15 113417      West     Arizona    AZ
 4       2110   3378        1.9      70.66   10.1      39.9    65  51945     South    Arkansas    AR
 5      21198   5114        1.1      71.71   10.3      62.6    20 156361      West  California    CA
 6       2541   4884        0.7      72.06    6.8      63.9   166 103766      West    Colorado    CO
 7       3100   5348        1.1      72.48    3.1      56.0   139   4862 Northeast Connecticut    CT
 8        579   4809        0.9      70.06    6.2      54.6   103   1982     South    Delaware    DE
 9       8277   4815        1.3      70.66   10.7      52.6    11  54090     South     Florida    FL
10       4931   4091        2.0      68.54   13.9      40.6    60  58073     South     Georgia    GA
# ... with 40 more rows
```

```
> summary(state)
   Population        Income       Illiteracy      Life.Exp         Murder
 Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96   Min.   : 1.400
 1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12   1st Qu.: 4.350
 Median : 2838   Median :4519   Median :0.950   Median :70.67   Median : 6.850
 Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88   Mean   : 7.378
 3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89   3rd Qu.:10.675
 Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60   Max.   :15.100
    HS.Grad          Frost             Area           State
 Min.   :37.80   Min.   :  0.00   Min.   :  1049   Length:50
 1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985   Class :character
 Median :53.25   Median :114.50   Median : 54277   Mode  :character
 Mean   :53.11   Mean   :104.46   Mean   : 70736
 3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81162
 Max.   :67.30   Max.   :188.00   Max.   :566432
```

```
> describe(state)
           vars  n     mean       sd   median  trimmed      mad     min      max     range  skew kurtosis       se
Population    1 50  4246.42  4464.49  2838.50  3384.28  2890.33  365.00 21198.0  20833.00  1.92     3.75   631.37
Income        2 50  4435.80   614.47  4519.00  4430.07   581.18 3098.00  6315.0   3217.00  0.20     0.24    86.90
Illiteracy    3 50     1.17     0.61     0.95     1.10     0.52   0.50     2.8      2.30  0.82    -0.47     0.09
Life.Exp      4 50    70.88     1.34    70.67    70.92     1.54  67.96    73.6      5.64 -0.15    -0.67     0.19
Murder        5 50     7.38     3.69     6.85     7.30     5.19   1.40    15.1     13.70  0.13    -1.21     0.52
HS.Grad       6 50    53.11     8.08    53.25    53.34     8.60  37.80    67.3     29.50 -0.32    -0.88     1.14
Frost         7 50   104.46    51.98   114.50   106.80    53.37   0.00   188.0    188.00 -0.37    -0.94     7.35
Area          8 50 70735.88 85327.30 54277.00 56575.72 35144.29 1049.00 566432.0 565383.00  4.10    20.39 12067.10
State*        9 50      NaN       NA       NA      NaN       NA    Inf     -Inf      -Inf    NA       NA       NA
```

```
 Descriptive statistics by group
group: Alabama
           vars n      mean sd   median  trimmed mad      min      max range skew kurtosis se
Population    1 1   3615.00 NA  3615.00  3615.00   0  3615.00  3615.00     0   NA       NA NA
Income        2 1   3624.00 NA  3624.00  3624.00   0  3624.00  3624.00     0   NA       NA NA
Illiteracy    3 1      2.10 NA     2.10     2.10   0     2.10     2.10     0   NA       NA NA
Life.Exp      4 1     69.05 NA    69.05    69.05   0    69.05    69.05     0   NA       NA NA
Murder        5 1     15.10 NA    15.10    15.10   0    15.10    15.10     0   NA       NA NA
HS.Grad       6 1     41.30 NA    41.30    41.30   0    41.30    41.30     0   NA       NA NA
Frost         7 1     20.00 NA    20.00    20.00   0    20.00    20.00     0   NA       NA NA
Area          8 1  50708.00 NA 50708.00 50708.00   0 50708.00 50708.00     0   NA       NA NA
State*        9 1       NaN NA       NA      NaN  NA      Inf     -Inf  -Inf   NA       NA NA
------------------------------------------------------------------------------------------------
group: Alaska
           vars n      mean sd   median  trimmed mad      min      max range skew kurtosis se
Population    1 1    365.00 NA   365.00   365.00   0   365.00   365.00     0   NA       NA NA
Income        2 1   6315.00 NA  6315.00  6315.00   0  6315.00  6315.00     0   NA       NA NA
```

```
> stat.desc(state)
              Population       Income Illiteracy     Life.Exp      Murder     HS.Grad        Frost         Area State
nbr.val     5.000000e+01 5.000000e+01 50.0000000 5.000000e+01  50.0000000  50.0000000   50.0000000 5.000000e+01    NA
nbr.null    0.000000e+00 0.000000e+00  0.0000000 0.000000e+00   0.0000000   0.0000000    1.0000000 0.000000e+00    NA
nbr.na      0.000000e+00 0.000000e+00  0.0000000 0.000000e+00   0.0000000   0.0000000    0.0000000 0.000000e+00    NA
min         3.650000e+02 3.098000e+03  0.5000000 6.796000e+01   1.4000000  37.8000000    0.0000000 1.049000e+03    NA
max         2.119800e+04 6.315000e+03  2.8000000 7.360000e+01  15.1000000  67.3000000  188.0000000 5.664320e+05    NA
range       2.083300e+04 3.217000e+03  2.3000000 5.640000e+00  13.7000000  29.5000000  188.0000000 5.653830e+05    NA
sum         2.123210e+05 2.217900e+05 58.5000000 3.543930e+03 368.9000000 2655.4000000 5223.0000000 3.536794e+06    NA
median      2.838500e+03 4.519000e+03  0.9500000 7.067500e+01   6.8500000  53.2500000  114.5000000 5.427700e+04    NA
mean        4.246420e+03 4.435800e+03  1.1700000 7.087860e+01   7.3780000  53.1080000  104.4600000 7.073588e+04    NA
SE.mean     6.313744e+02 8.689917e+01  0.0862010 1.898431e-01   0.5220626   1.1422600    7.3512020 1.206710e+04    NA
CI.mean.0.95 1.268794e+03 1.746304e+02  0.1732274 3.815040e-01   1.0491240   2.2954574   14.7727936 2.424975e+04    NA
var         1.993168e+07 3.775733e+05  0.3715306 1.802020e+00  13.6274653  65.2378939 2702.0085714 7.280748e+09    NA
std.dev     4.464491e+03 6.144699e+02  0.6095331 1.342394e+00   3.6915397   8.0769978   51.9808481 8.532730e+04    NA
coef.var    1.051354e+00 1.385252e-01  0.5209685 1.893934e-02   0.5003442   0.1520863    0.4976149 1.206280e+00    NA
```

Part 1

Data report overview

The dataset examined has the following dimensions:

Feature	Result
Number of observations	50
Number of variables	9

Checks performed

The following variable checks were performed, depending on the data type of each variable:

	character	factor	labelled	numeric	integer	logical	Date
Identify miscoded missing values	×	×	×	×	×		×
Identify prefixed and suffixed whitespace	×	×	×				
Identify levels with < 6 obs.	×	×	×				
Identify case issues	×	×	×				
Identify misclassified numeric or integer variables	×	×	×				
Identify outliers				×	×		×

Please note that all numerical values in the following have been rounded to 2 decimals.

Part 2

Summary table

	Variable class	# unique values	Missing observations	Any problems?
Population	numeric	50	0.00 %	×
Income	numeric	50	0.00 %	×
Illiteracy	numeric	20	0.00 %	
Life.Exp	numeric	47	0.00 %	×
Murder	numeric	44	0.00 %	
HS.Grad	numeric	47	0.00 %	
Frost	numeric	43	0.00 %	×
Area	numeric	50	0.00 %	×
State	character	50	0.00 %	×

```
> diff_data(state,state2)
Daff Comparison: 'state' vs. 'state2'
  First 6 and last 6 patch lines:
     @:@          A:A     B:B        C:C      D:D    E:E      F:F   G:G    H:H
1      @@   Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
2    1:1  ->  3615->3616   3624        2.1    69.05   15.1    41.3    20  50708
3    2:2  ->    365->366   6315        1.5    69.31   11.3    66.7   152 566432
4    3:3  ->  2212->2213   4530        1.8    70.55    7.8    58.1    15 113417
5    4:4  ->  2110->2111   3378        1.9    70.66   10.1    39.9    65  51945
6    5:5  -> 21198->21199   5114        1.1    71.71   10.3    62.6    20 156361
...  ... ...         ...    ...        ...      ...    ...     ...   ...    ...
47 45:45  ->    472->473   3907        0.6    71.64    5.5    57.1   168   9267
48 46:46  ->  4981->4982   4701        1.4    70.08    9.5    47.8    85  39780
49 47:47  ->  3559->3560   4864        0.6    71.72    4.3    63.5    32  66570
50 48:48  ->  1799->1800   3617        1.4    69.48    6.7    41.6   100  24070
51 49:49  ->  4589->4590   4468        0.7    72.48      3    54.5   149  54464
52 50:50  ->    376->377   4566        0.6    70.29    6.9    62.9   173  97203
```

‘state’ vs. ‘state2’

2018-04-01 19:05:29

	#	Modified	Reordered	Deleted	Added
Rows	50	49	1	1	1
Columns	8	0	0	0	0

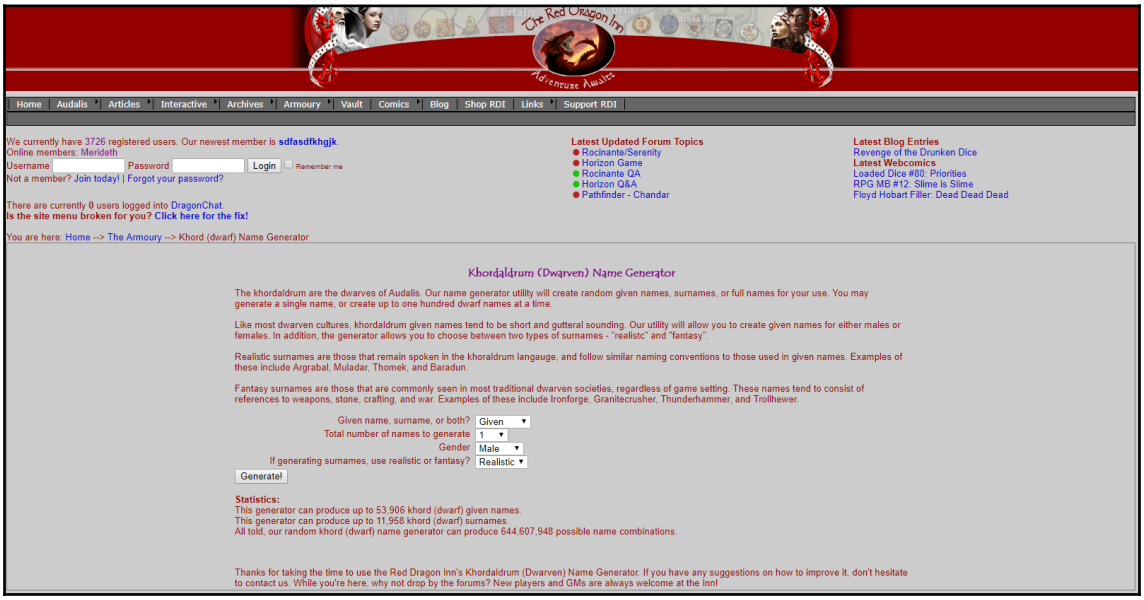
@:@		A:A	B:B	C:C	D:D	E:E	F:F	G:G	H:H
	@@	**Population**	**Income**	**Illiteracy**	**Life.Exp**	**Murder**	**HS.Grad**	**Frost**	**Area**
1:1	⇒	3615 → 3616	3624	2.1	69.05	15.1	41.3	20	50708
2:2	⇒	365 → 366	6315	1.5	69.31	11.3	66.7	152	566432
3:3	⇒	2212 → 2213	4530	1.8	70.55	7.8	58.1	15	113417
4:4	⇒	2110 → 2111	3378	1.9	70.66	10.1	39.9	65	51945
5:5	⇒	21198 → 21199	5114	1.1	71.71	10.3	62.6	20	156361
6:6	⇒	2541 → 2542	4884	0.7	72.06	6.8	63.9	166	103766
7:7	⇒	3100 → 3101	5348	1.1	72.48	3.1	56	139	4862
8:8	⇒	579 → 580	4809	0.9	70.06	6.2	54.6	103	1982
9:9	⇒	8277 → 8278	4815	1.3	70.66	10.7	52.6	11	54090
10:10	⇒	4931 → 4932	4091	2	68.54	13.9	40.6	60	58073
11:11	⇒	868 → 869	4963	1.9	73.6	6.2	61.9	0	6425
-:12	+++	814	4119	0.6	71.87	5.3	59.5	126	82677
13:13	⇒	11197 → 11198	5107	0.9	70.14	10.3	52.6	127	55748
14:14	⇒	5313 → 5314	4458	0.7	70.88	7.1	52.9	122	36097
15:15	⇒	2861 → 2862	4628	0.5	72.56	2.3	59	140	55941
16:16	⇒	2280 → 2281	4669	0.6	72.58	4.5	59.9	114	81787
17:17	⇒	3387 → 3388	3712	1.6	70.1	10.6	38.5	95	39650

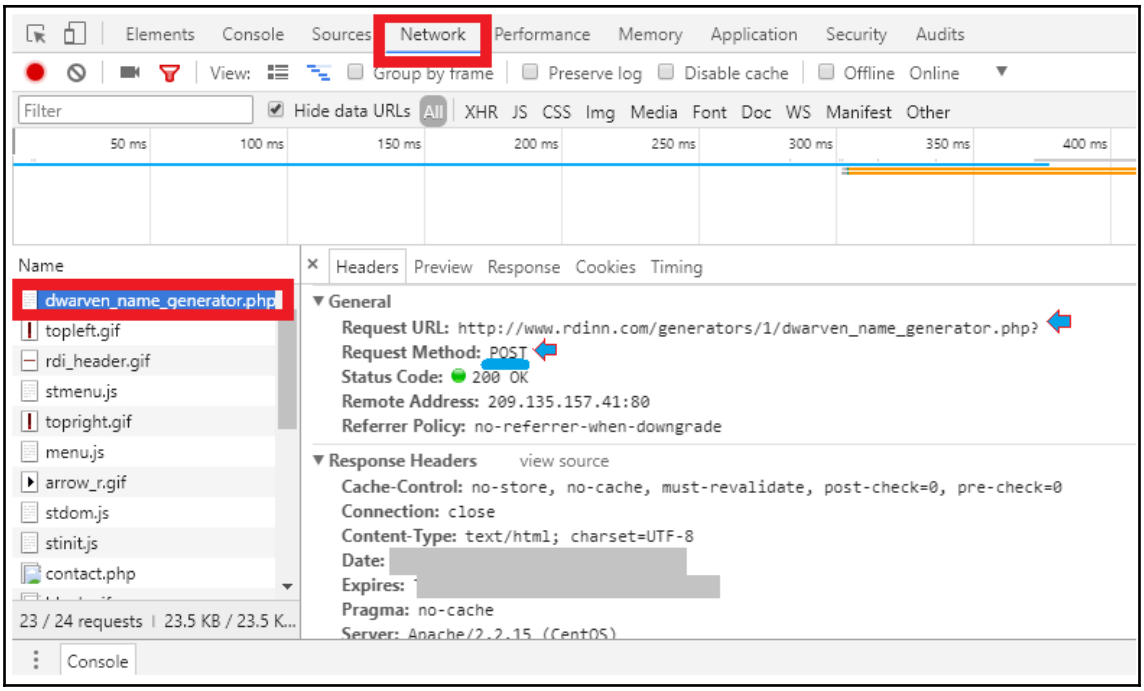
```
> describe(flights)
               vars      n    mean       sd median trimmed     mad  min  max range  skew kurtosis   se
year              1 336776 2013.00     0.00   2013 2013.00    0.00 2013 2013     0   NaN      NaN 0.00
month             2 336776    6.55     3.41      7    6.56    4.45    1   12    11 -0.01    -1.19 0.01
day               3 336776   15.71     8.77     16   15.70   11.86    1   31    30  0.01    -1.19 0.02
dep_time          4 328521 1349.11   488.28   1401 1346.82  634.55    1 2400  2399 -0.02    -1.09 0.85
sched_dep_time    5 336776 1344.25   467.34   1359 1341.60  613.80  106 2359  2253 -0.01    -1.20 0.81
dep_delay         6 328521   12.64    40.21     -2    3.32    5.93  -43 1301  1344  4.80    43.95 0.07
arr_time          7 328063 1502.05   533.26   1535 1526.42  619.73    1 2400  2399 -0.47    -0.19 0.93
sched_arr_time    8 336776 1536.38   497.46   1556 1550.67  618.24    1 2359  2358 -0.35    -0.38 0.86
arr_delay         9 327346    6.90    44.63     -5   -1.03   20.76  -86 1272  1358  3.72    29.23 0.08
carrier*         10 336776    9.00     0.00      9    9.00    0.00    9    9     0   NaN      NaN 0.00
flight           11 336776 1971.92  1632.47   1496 1830.51 1608.62    1 8500  8499  0.66    -0.85 2.81
tailnum*         12 334264     NaN       NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
origin*          13 336776     NaN       NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
dest*            14 336776     NaN       NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
air_time         15 327346  150.69    93.69    129  140.03   75.61   20  695   675  1.07     0.86 0.16
distance         16 336776 1039.91   733.23    872  955.27  569.32   17 4983  4966  1.13     1.19 1.26
hour             17 336776   13.18     4.66     13   13.15    5.93    1   23    22  0.00    -1.21 0.01
minute           18 336776   26.23    19.30     29   25.64   23.72    0   59    59  0.09    -1.24 0.03
time_hour*       19 336776     NaN       NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
```

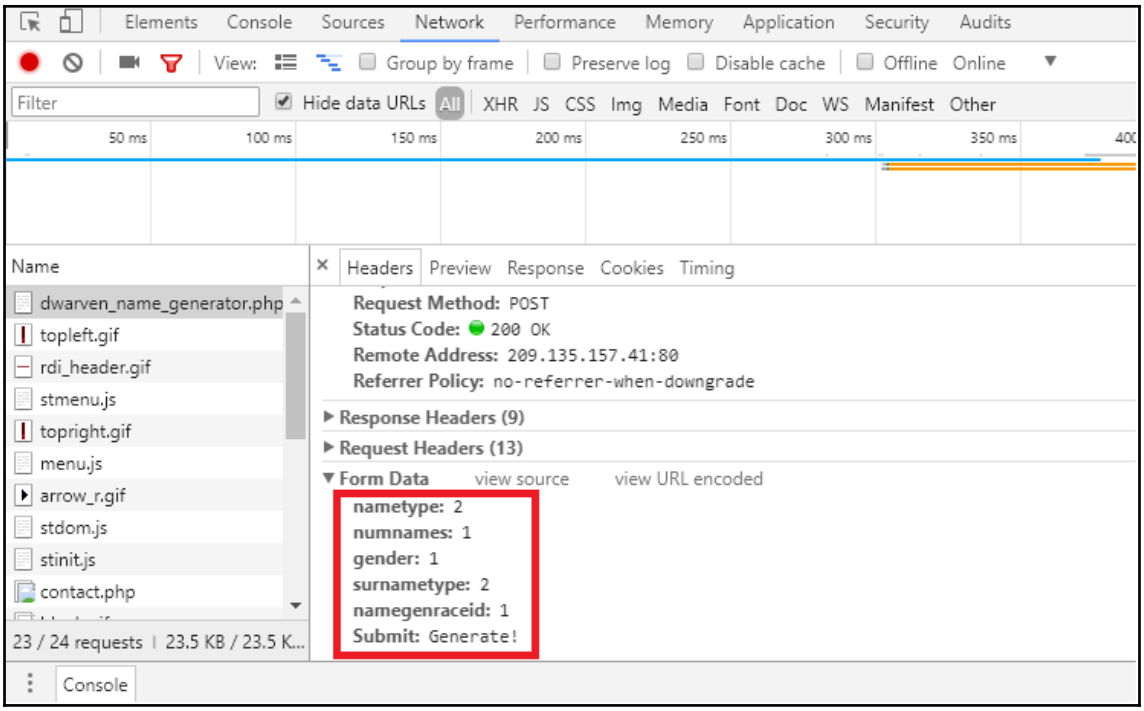
```
 Descriptive statistics by group
group: EWR
               vars      n    mean      sd median trimmed     mad  min  max range  skew kurtosis   se
year              1 120835 2013.00    0.00   2013 2013.00    0.00 2013 2013     0   NaN      NaN 0.00
month             2 120835    6.49    3.42      6    6.49    4.45    1   12    11  0.01    -1.19 0.01
day               3 120835   15.70    8.76     16   15.68   11.86    1   31    30  0.01    -1.18 0.03
dep_time          4 117596 1336.70  487.01   1341 1331.47  618.24    1 2400  2399  0.02    -1.14 1.42
sched_dep_time    5 120835 1322.47  465.37   1330 1318.27  594.52  106 2345  2239  0.01    -1.21 1.34
dep_delay         6 117596   15.11   41.32     -1    5.51    7.41  -25 1126  1151  4.15    30.97 0.12
arr_time          7 117445 1491.88  529.05   1522 1510.27  612.31    1 2400  2399 -0.40    -0.27 1.54
sched_arr_time    8 120835 1527.98  486.89   1542 1535.15  596.01    1 2359  2358 -0.22    -0.63 1.40
arr_delay         9 117127    9.11   45.53     -4    0.83   20.76  -86 1109  1195  3.35    21.99 0.13
carrier*         10 120835    9.00    0.00      9    9.00    0.00    9    9     0   NaN      NaN 0.00
flight           11 120835 2373.51 1746.61   1637 2309.54 1879.94    1 6181  6180  0.32    -1.45 5.02
tailnum*         12 120229     NaN      NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
origin*          13 120835     NaN      NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
dest*            14 120835     NaN      NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
air_time         15 117127  153.30   93.34    130  143.20   71.16   20  695   675  1.11     1.18 0.27
distance         16 120835 1056.74  730.22    872  974.42  508.53   17 4963  4946  1.23     1.78 2.10
hour             17 120835   12.95    4.65     13   12.91    5.93    1   23    22  0.02    -1.22 0.01
minute           18 120835   27.24   18.15     29   26.89   22.24    0   59    59  0.07    -1.07 0.05
time_hour*       19 120835     NaN      NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
------------------------------------------------------------------------------------------------------
group: JFK
               vars      n    mean      sd median trimmed     mad  min  max range  skew kurtosis   se
year              1 111279 2013.00    0.00   2013 2013.00    0.00 2013 2013     0   NaN      NaN 0.00
month             2 111279    6.50    3.41      7    6.50    4.45    1   12    11  0.00    -1.18 0.01
day               3 111279   15.73    8.79     16   15.73   11.86    1   31    30  0.00    -1.19 0.03
dep_time          4 109416 1398.57  505.53   1500 1403.10  630.11    1 2400  2399 -0.17    -1.02 1.53
sched_dep_time    5 111279 1401.93  482.27   1459 1403.55  636.04  540 2359  1819 -0.11    -1.20 1.45
dep_delay         6 109416   12.11   39.04     -1    3.07    5.93  -43 1301  1344  5.45    64.43 0.12
arr_time          7 109284 1520.07  579.09   1625 1565.91  690.89    1 2400  2399 -0.66    -0.08 1.75
sched_arr_time    8 111279 1564.98  544.69   1647 1599.73  641.97    1 2359  2358 -0.64     0.00 1.63
arr_delay         9 109079    5.55   44.28     -6   -2.03   20.76  -79 1272  1351  3.99    38.99 0.13
carrier*         10 111279    9.00    0.00      9    9.00    0.00    9    9     0   NaN      NaN 0.00
flight           11 111279 1365.75 1376.74    801 1181.75 1009.65    1 5765  5764  1.05     0.10 4.13
tailnum*         12 110370     NaN      NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
origin*          13 111279     NaN      NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
dest*            14 111279     NaN      NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
air_time         15 109079  178.35  113.79    149  172.81  139.36   21  691   670  0.49    -0.75 0.34
distance         16 111279 1266.25  896.11   1069 1229.73 1138.64   94 4983  4889  0.48    -0.67 2.69
hour             17 111279   13.74    4.80     14   13.77    5.93    5   23    18 -0.11    -1.22 0.01
minute           18 111279   27.50   19.36     30   27.24   22.24    0   59    59  0.00    -1.26 0.06
time_hour*       19 111279     NaN      NA     NA     NaN      NA  Inf -Inf  -Inf    NA       NA   NA
------------------------------------------------------------------------------------------------------
group: LGA
               vars      n    mean      sd median trimmed     mad  min  max range  skew kurtosis   se
```

```
> flights3[order(-AvgTime)]
     dest origin MaxTime MinTime   AvgTime                                 dest_name
  1:  HNL    JFK     691     580 623.08772           Daniel K Inouye International Airport
  2:  HNL    EWR     695     562 612.07521           Daniel K Inouye International Airport
  3:  ANC    EWR     434     388 413.12500     Ted Stevens Anchorage International Airport
  4:  SFO    JFK     490     301 347.40363             San Francisco International Airport
  5:  SJC    JFK     396     305 346.60671 Norman Y. Mineta San Jose International Airport
 ---
225:  PHL    JFK      61      21  30.83687                                     Erase Me 19
226:  PHL    EWR      39      21  28.66667             Philadelphia International Airport
227:  PHL    EWR      39      21  28.66667                                     Erase Me 19
228:  BDL    EWR      56      20  25.46602                  Bradley International Airport
229:  LGA    EWR    -Inf     Inf       NaN                             La Guardia Airport
```

Chapter 4: KDD, Data Mining, and Text Mining



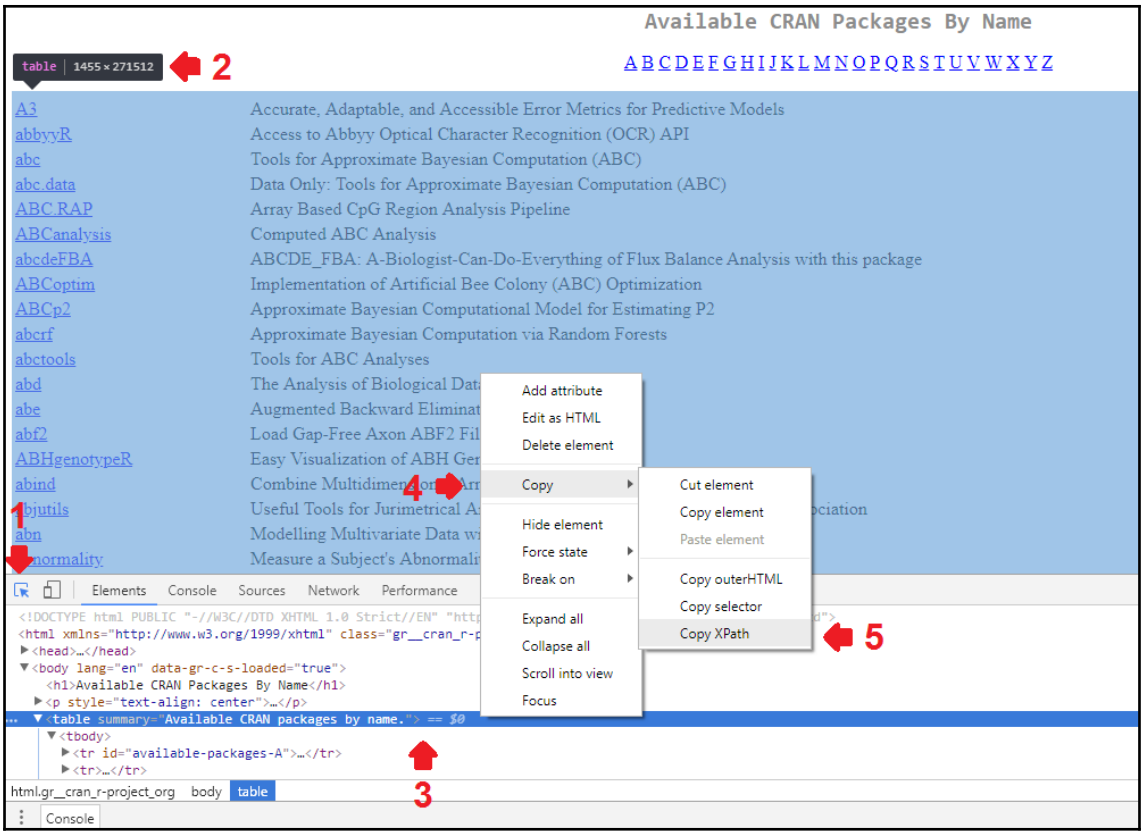

Elements
Console
Sources
Network
Performance
Memory
Application
Security
Audits
View:
Group by frame
Preserve log
Disable cache
Offline
Online
Filter
Hide data URLs
All
XHR
JS
CSS
Img
Media
Font
Doc
WS
Manifest
Other
50 ms
100 ms
150 ms
200 ms
250 ms
300 ms
350 ms
400 ms
Name
dwarven_name_generator.php
topleft.gif
rdi_header.gif
stmenu.js
topright.gif
menu.js
arrow_r.gif
stdom.js
stinit.js
contact.php
23 / 24 requests
23.5 KB / 23.5 K...
Headers
Preview
Response
Cookies
Timing
General
Request URL: http://www.rdinn.com/generators/1/dwarven_name_generator.php?
Request Method: POST
Status Code: 200 OK
Remote Address: 209.135.157.41:80
Referrer Policy: no-referrer-when-downgrade
Response Headers
view source
Cache-Control: no-store, no-cache, must-revalidate, post-check=0, pre-check=0
Connection: close
Content-Type: text/html; charset=UTF-8
Date:
Expires:
Pragma: no-cache
Server: Apache/2.2.15 (CentOS)
Console

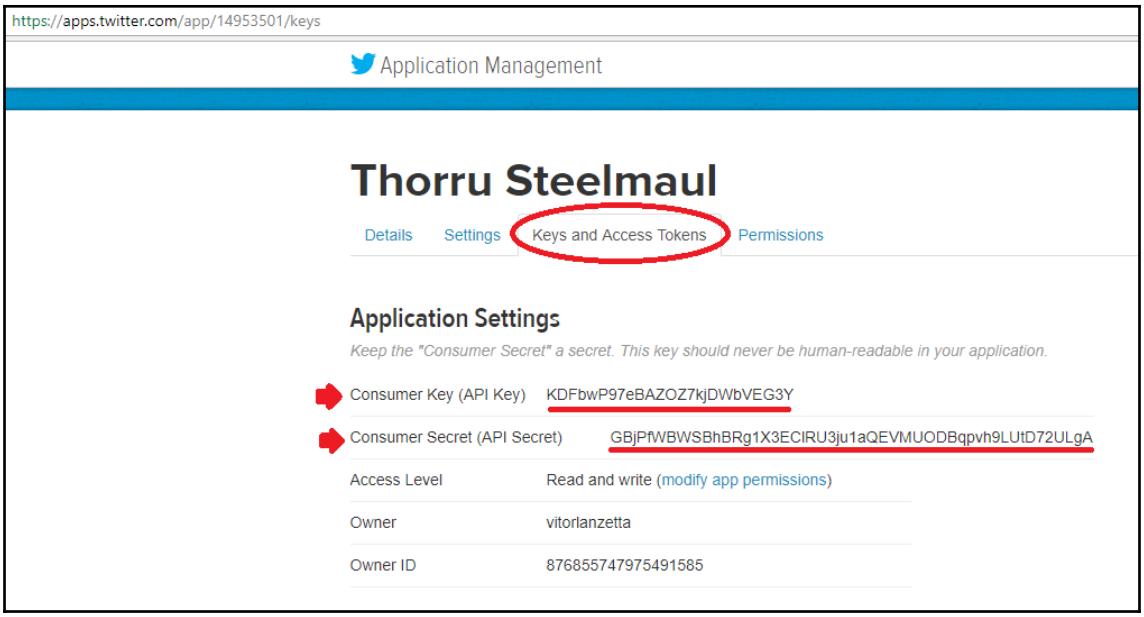

Elements
Console
Sources
Network
Performance
Memory
Application
Security
Audits
View:
Group by frame
Preserve log
Disable cache
Offline
Online
Filter
Hide data URLs
All
XHR
JS
CSS
Img
Media
Font
Doc
WS
Manifest
Other
50 ms
100 ms
150 ms
200 ms
250 ms
300 ms
350 ms
Name
dwarven_name_generator.php
topleft.gif
rdi_header.gif
stmenu.js
topright.gif
menu.js
arrow_r.gif
stdom.js
stinit.js
contact.php
23 / 24 requests
23.5 KB / 23.5 K...
Headers
Preview
Response
Cookies
Timing
Request Method: POST
Status Code: 200 OK
Remote Address: 209.135.157.41:80
Referrer Policy: no-referrer-when-downgrade
Response Headers (9)
Request Headers (13)
Form Data
view source
view URL encoded
nametype: 2
numnames: 1
gender: 1
surnametype: 2
namegenraceid: 1
Submit: Generate!
Console

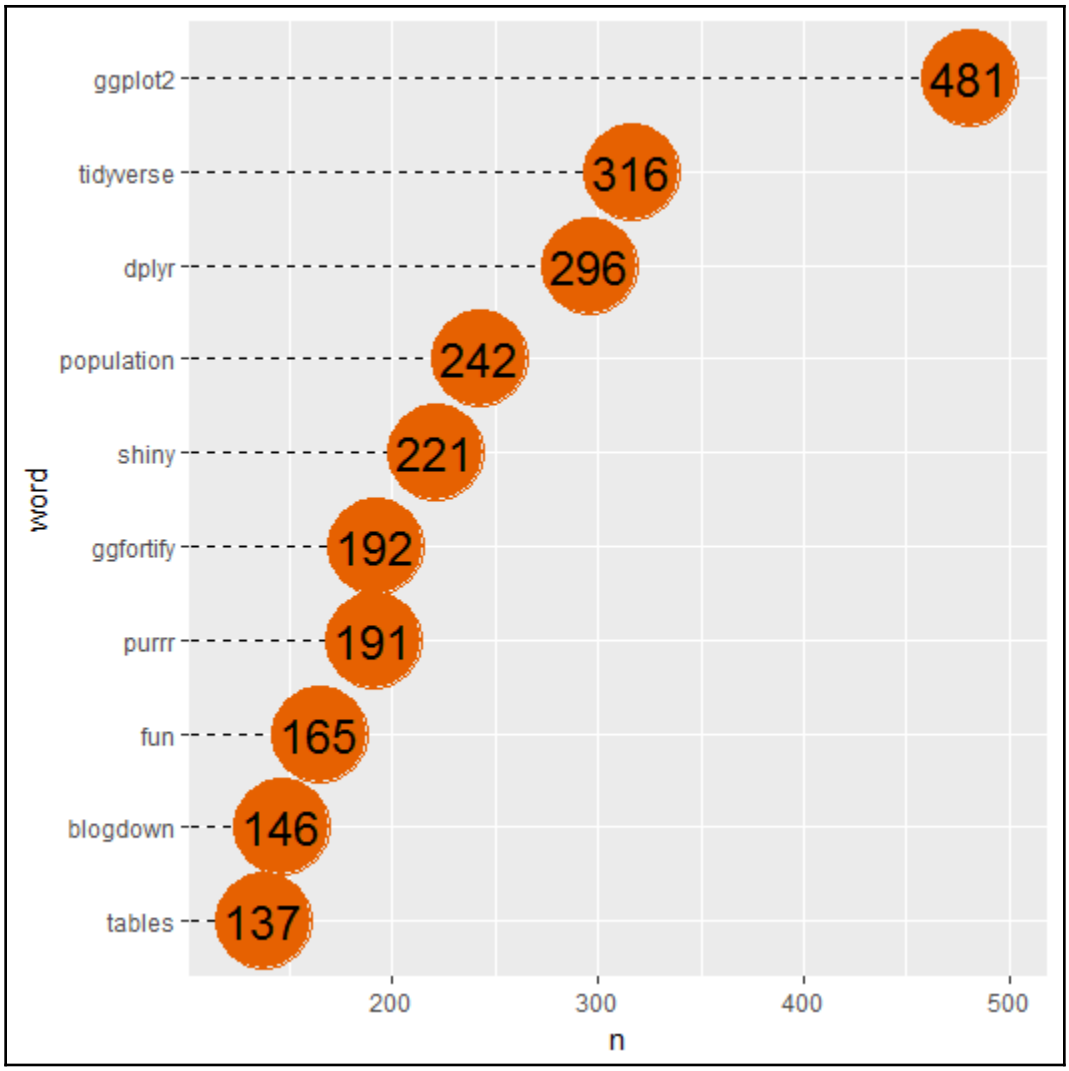
Available CRAN Packages By Name

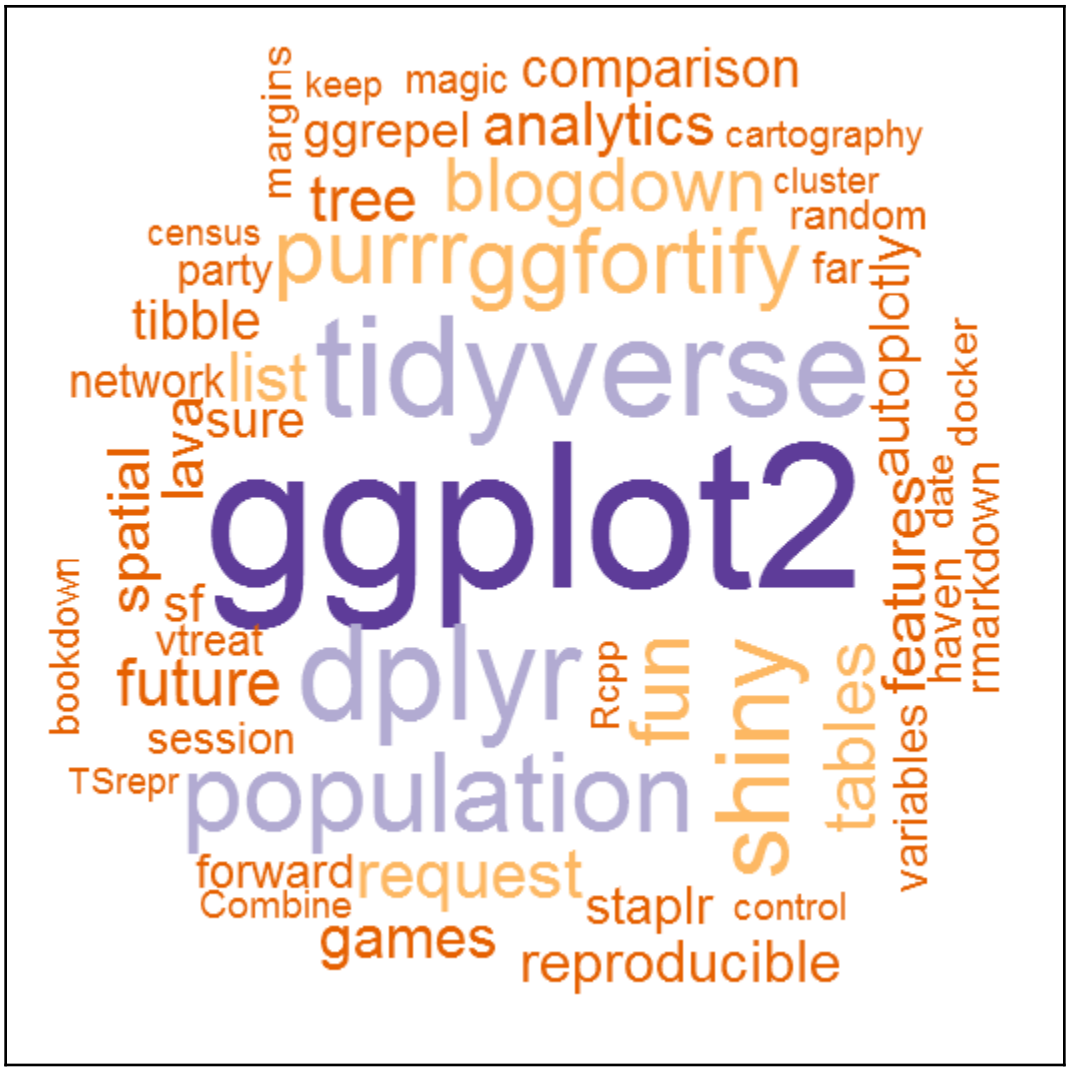
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

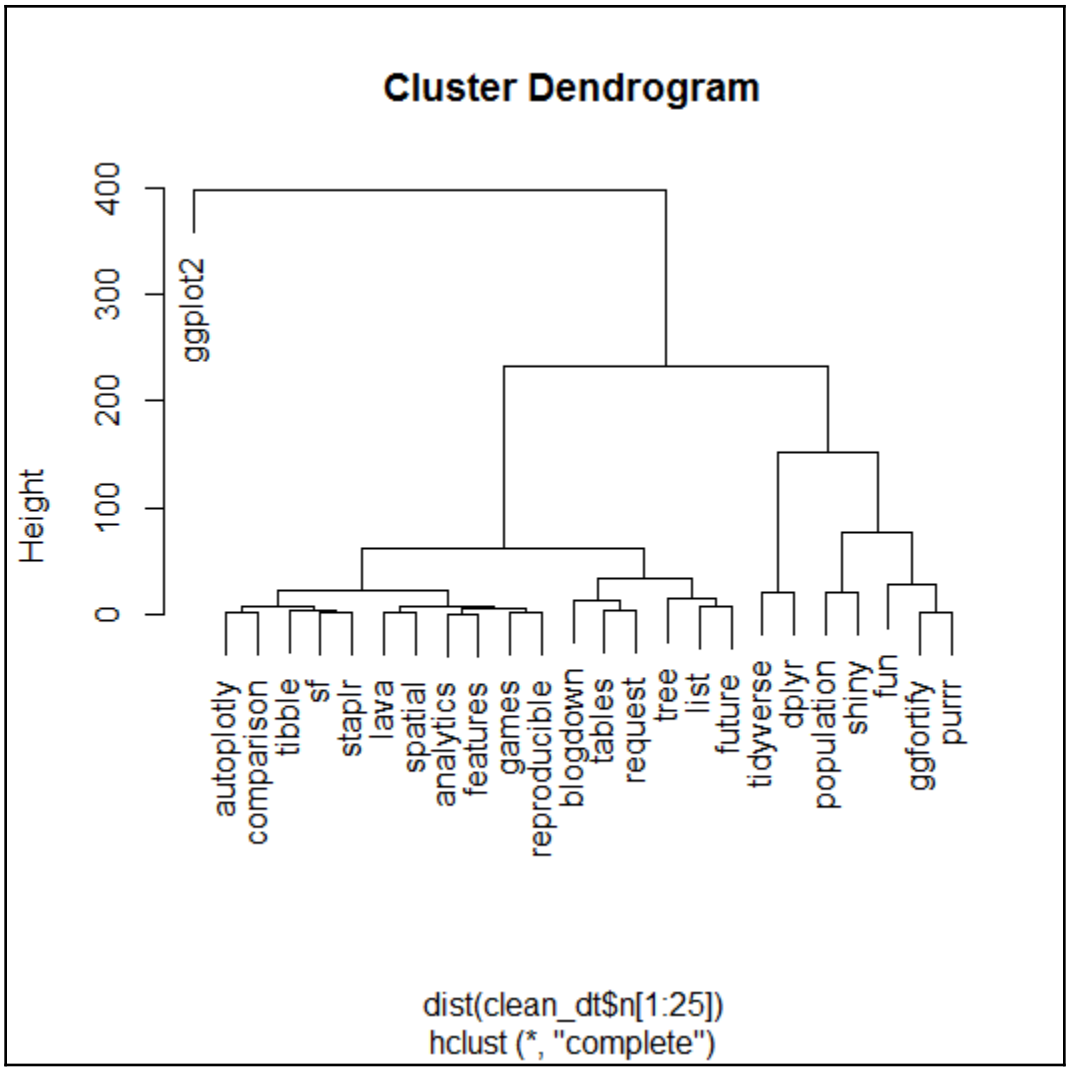
A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
ABC.RAP	Array Based CpG Region Analysis Pipeline
ABCanalysis	Computed ABC Analysis
abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization
ABCp2	Approximate Bayesian Computational Model for Estimating P2
abcrf	Approximate Bayesian Computation via Random Forests
abctools	Tools for ABC Analyses
abd	The Analysis of Biological Data
abe	Augmented Backward Elimination
abf2	Load Gap-Free Axon ABF2 Files
ABHgenotypeR	Easy Visualization of ABH Genotypes
abind	Combine Multidimensional Arrays
abjutils	Useful Tools for Jurimetrical Analysis Used by the Brazilian Jurimetrics Association
abn	Modelling Multivariate Data with Additive Bayesian Networks
abnormality	Measure a Subject's Abnormality with Respect to a Reference Population
abodOutlier	Angle-Based Outlier Detection

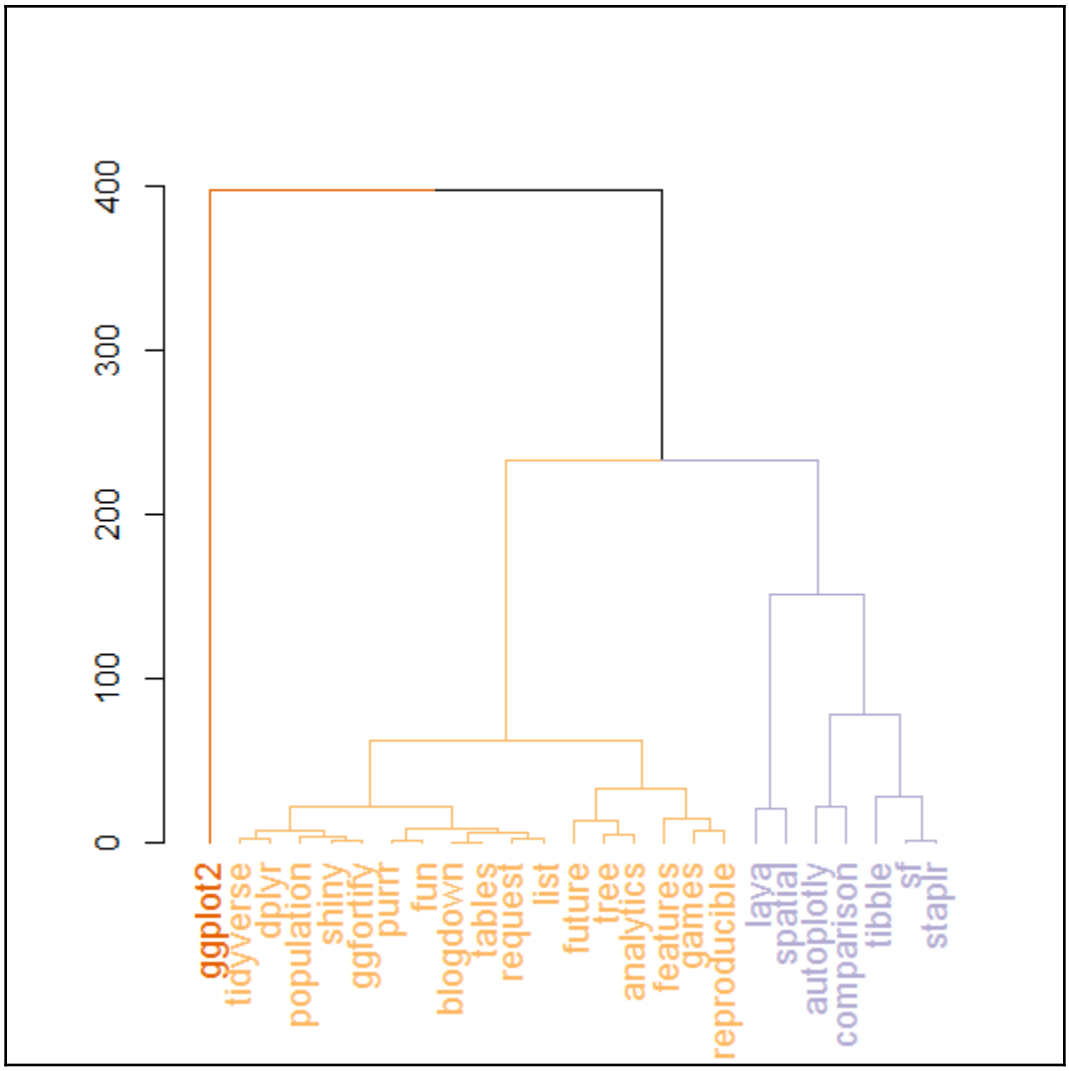

Available CRAN Packages By Name
table | 1455 × 271512
2
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
A3 Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
abbyyR Access to Abbyy Optical Character Recognition (OCR) API
abc Tools for Approximate Bayesian Computation (ABC)
abc.data Data Only: Tools for Approximate Bayesian Computation (ABC)
ABC.RAP Array Based CpG Region Analysis Pipeline
ABCanalysis Computed ABC Analysis
abcdeFBA ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCoptim Implementation of Artificial Bee Colony (ABC) Optimization
ABCp2 Approximate Bayesian Computational Model for Estimating P2
abcrf Approximate Bayesian Computation via Random Forests
abctools Tools for ABC Analyses
abd The Analysis of Biological Data
abe Augmented Backward Eliminat
abf2 Load Gap-Free Axon ABF2 Fil
ABHgenotypeR Easy Visualization of ABH Ge
abind Combine Multidimen
bjutils Useful Tools for Jurimetrical A
abn Modelling Multivariate Data wi
normality Measure a Subject's Abnormali
1
4
Add attribute
Edit as HTML
Delete element
Copy
Hide element
Force state
Break on
Expand all
Collapse all
Scroll into view
Focus
Cut element
Copy element
Paste element
Copy outerHTML
Copy selector
Copy XPath
5
Elements Console Sources Network Performance
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http
<html xmlns="http://www.w3.org/1999/xhtml" class="gr__cran_r-p
<head>…</head>
<body lang="en" data-gr-c-s-loaded="true">
<h1>Available CRAN Packages By Name</h1>
…
<table summary="Available CRAN packages by name."> == $0
<tbody>
<tr id="available-packages-A">…</tr>
<tr>…</tr>
3
html.gr__cran_r-project_org body table
Console


https://apps.twitter.com/app/14953501/keys
Application Management
Thorru Steelmaul
Details
Settings
Keys and Access Tokens
Permissions
Application Settings
Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.
Consumer Key (API Key)
KDFbwP97eBAZOZ7kjDWbVEG3Y
Consumer Secret (API Secret)
GBjPfWBWSBhBRg1X3ECIRU3ju1aQEVMUODBqpvh9LUtD72ULgA
Access Level
Read and write (modify app permissions)
Owner
vitorlanzetta
Owner ID
876855747975491585


ggplot2
tidyverse
dplyr
population
shiny
ggfortify
purrr
fun
blogdown
tables
481
316
296
242
221
192
191
165
146
137
word
200
300
400
500
n


margins
keep
magic
comparison
ggrepel
analytics
cartography
tree
blogdown
cluster
random
census
party
purrr
ggfortify
far
tibble
tidyverse
autoplotly
network
list
lava
sure
docker
ggplot2
spatial
features
date
haven
rmarkdown
bookdown
sf
vtreat
future
dplyr
Rcpp
fun
shiny
tables
variables
session
TSrepr
population
forward
request
Combine
staplr
control
games
reproducible


Cluster Dendrogram
Height
0
100
200
300
400
ggplot2
autoplotly
comparison
tibble
sf
staplr
lava
spatial
analytics
features
games
reproducible
blogdown
tables
request
tree
list
future
tidyverse
dplyr
population
shiny
fun
ggfortify
purrr
dist(clean_dt$n[1:25])
hclust (*, "complete")

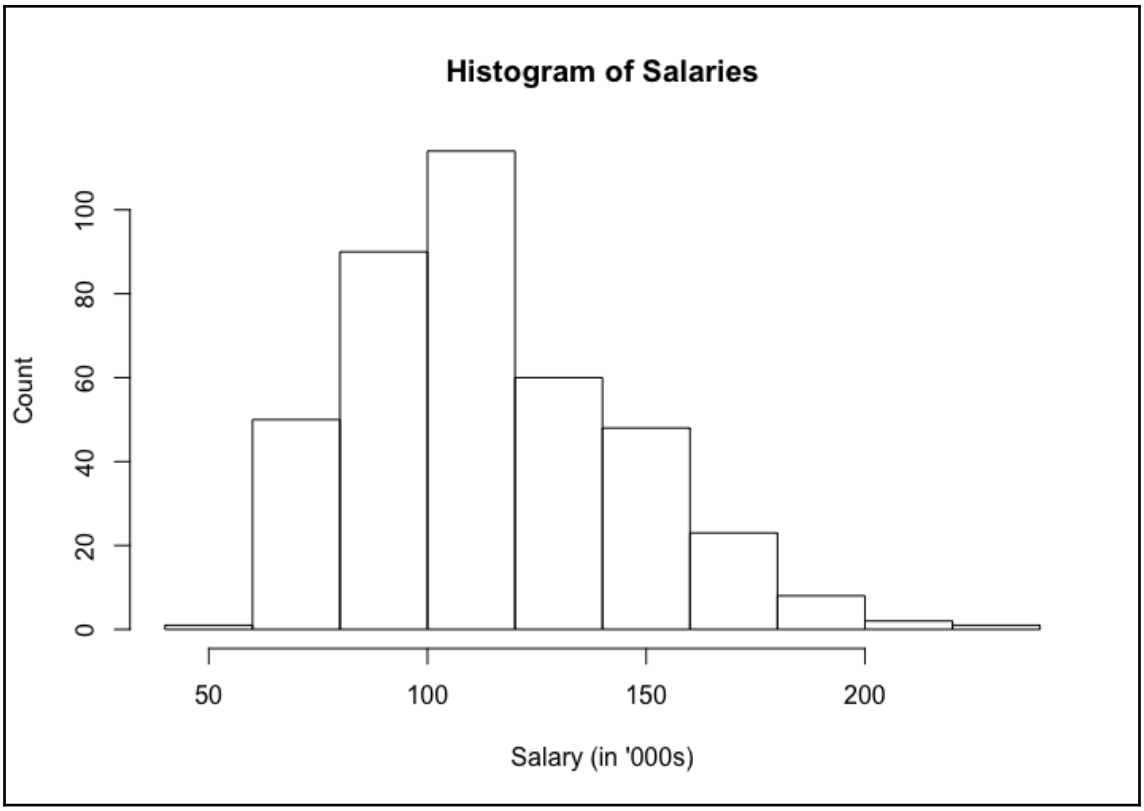

0
100
200
300
400
ggplot2
tidyverse
dplyr
population
shiny
ggfortify
purrr
fun
blogdown
tables
request
list
future
tree
analytics
features
games
reproducible
laya
spatial
autoplotly
comparison
tibble
sf
staplr

Chapter 5: Data Analysis with R

Concept	Terminology	Meaning
Type of Variable	Continuous/Quantitative	Numbers on which you can perform arithmetic
	Discrete/Categorical	Qualitative – Alphabetical, or Numeric that simply denotes a 'category'
Variable in an equation	Dependent	The Left-Hand-Side, the y in y = x + 1
	Independent	The Right-Hand-Side, every variable other than the y in y = x + 1 (i.e., x in this case)

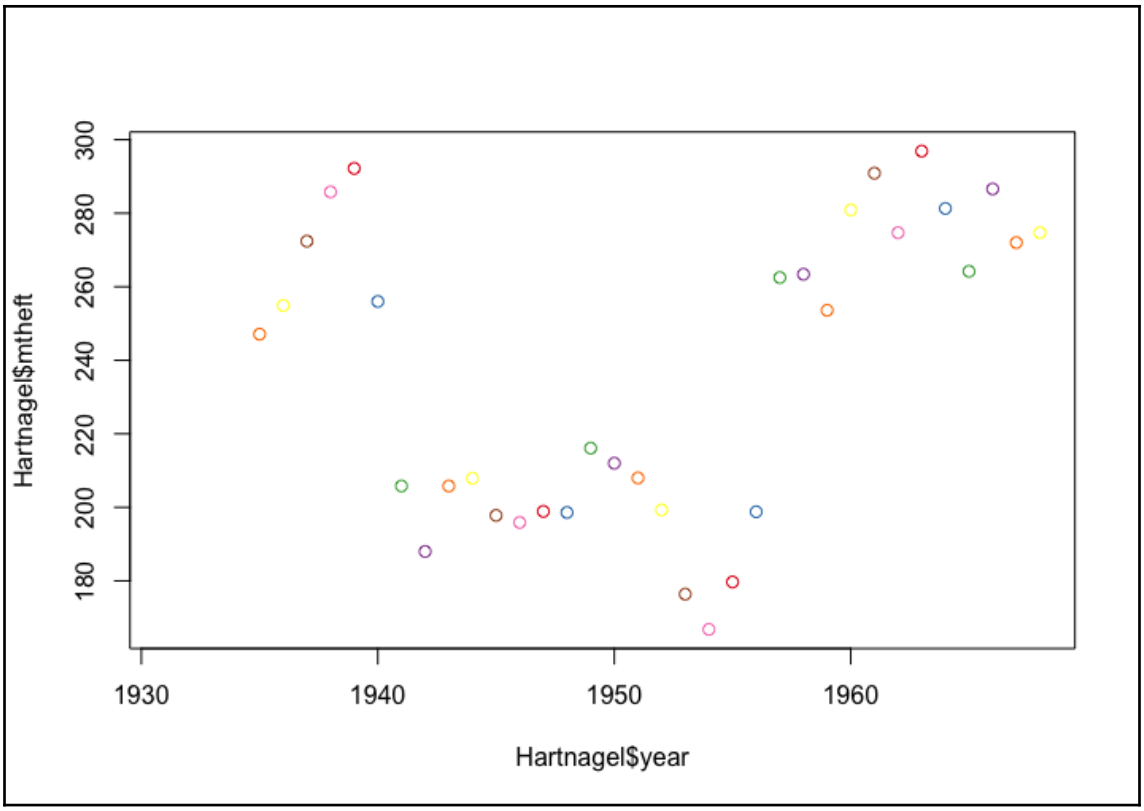
```
> head(car_data)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
>
```

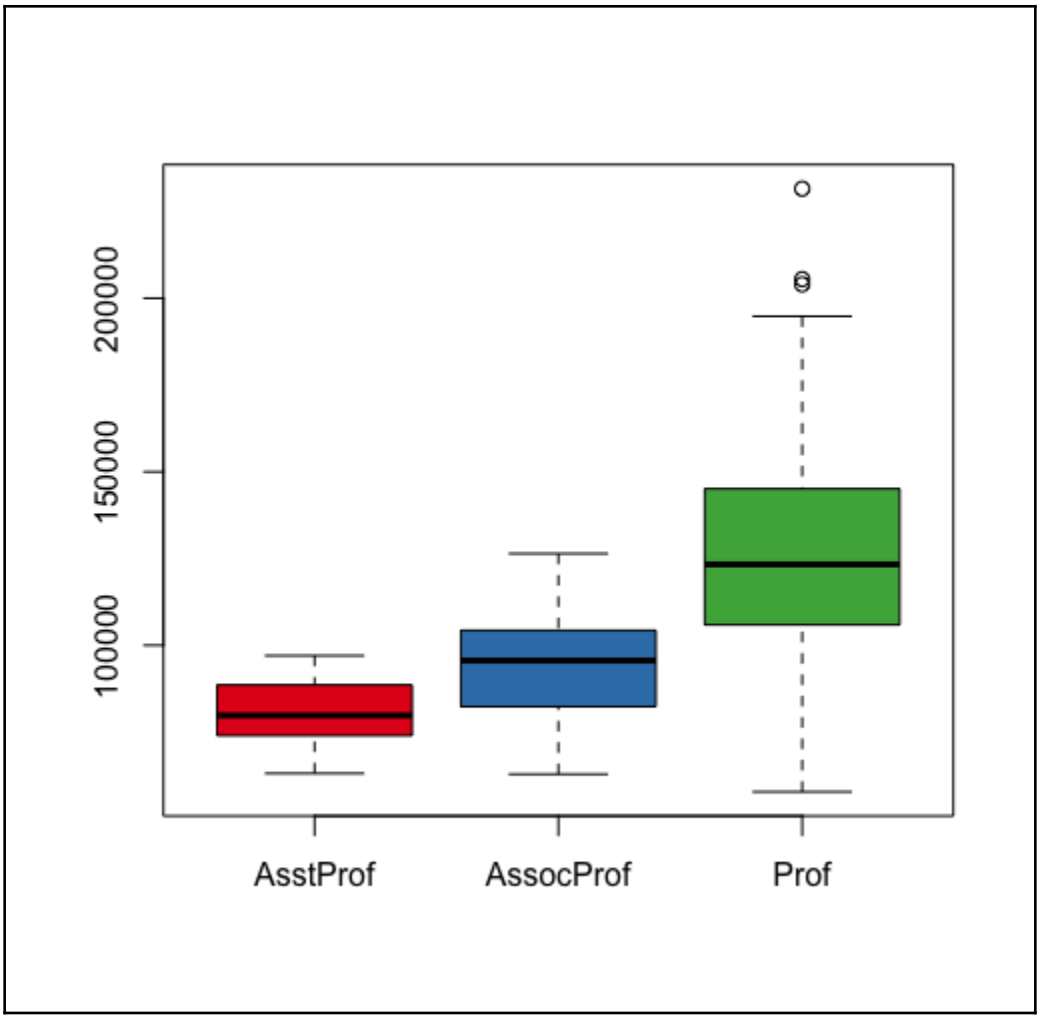
```
> glimpse(read_csv("car_data.csv")) # Tidyverse
Parsed with column specification:
cols(
  mpg = col_double(),
  cyl = col_integer(),
  disp = col_double(),
  hp = col_integer(),
  drat = col_double(),
  wt = col_double(),
  qsec = col_double(),
  vs = col_integer(),
  am = col_integer(),
  gear = col_integer(),
  carb = col_integer()
)
Observations: 32
Variables: 11
$ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2, 10.4, 10.4, 14.7, 32.4,...
$ cyl  <int> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8, 8, 8, 8, 4, 4, 4, 8, 6, 8, 4
$ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 167.6, 167.6, 275.8, 275.8, 275.8, 472.0, 46...
$ hp   <int> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180, 205, 215, 230, 66, 52, 65, 97, 150, 150,...
$ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92, 3.07, 3.07, 3.07, 2.93, 3.00, 3.23, 4.08,...
$ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3.440, 3.440, 4.070, 3.730, 3.780, 5.250, 5....
$ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 18.30, 18.90, 17.40, 17.60, 18.00, 17.98, 17...
$ vs   <int> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1
$ am   <int> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1
$ gear <int> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 4, 5, 5, 5, 5, 5, 4
$ carb <int> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2, 2, 4, 2, 1, 2, 2, 4, 6, 8, 2
```

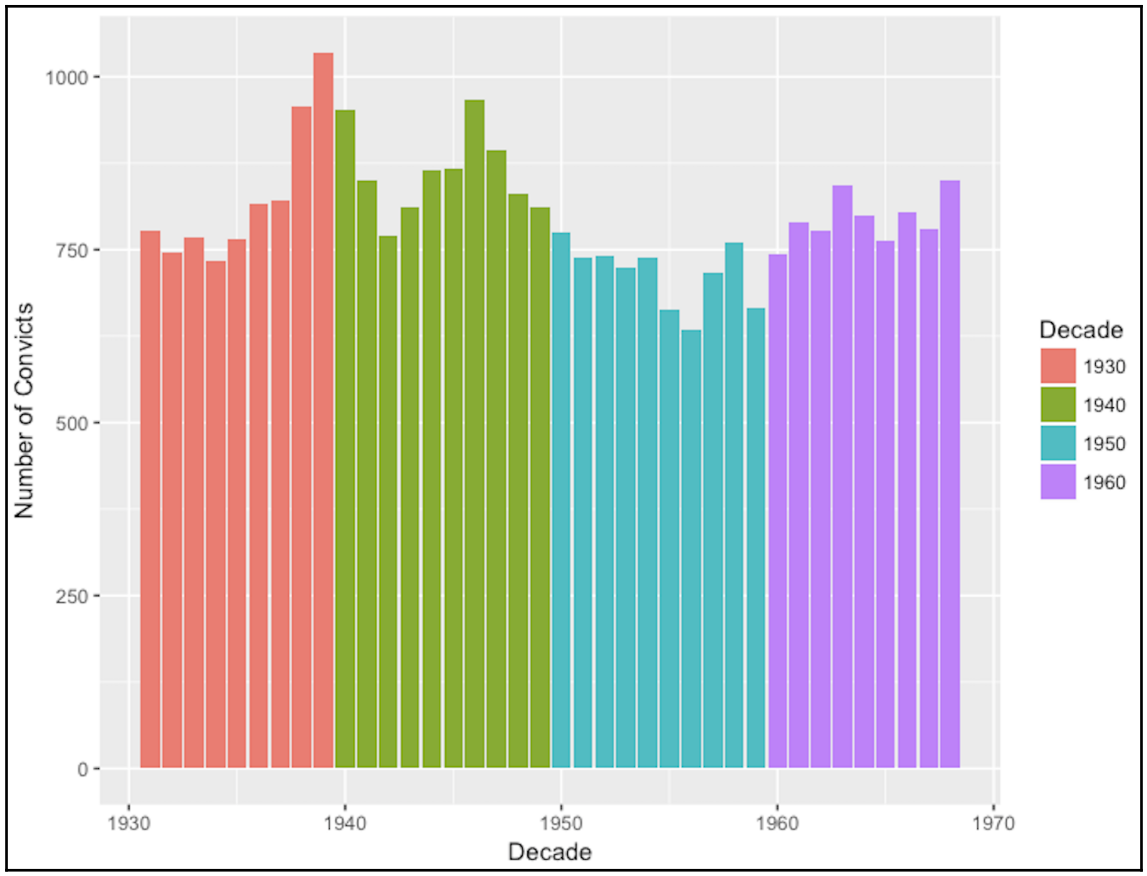

Histogram of Salaries
Count
0
20
40
60
80
100
50
100
150
200
Salary (in '000s)

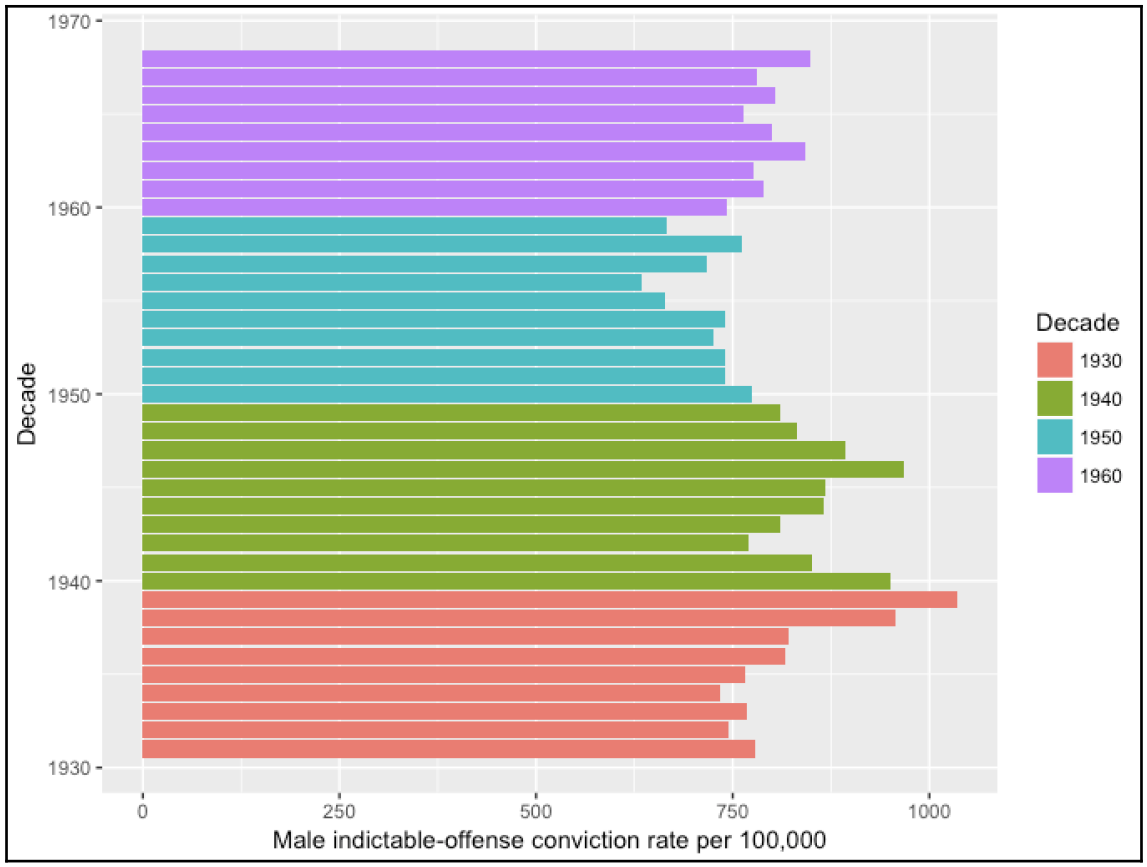

Histogram of Salaries
Count
0
20
40
60
80
100
50
100
150
200
Salary (in '000s)

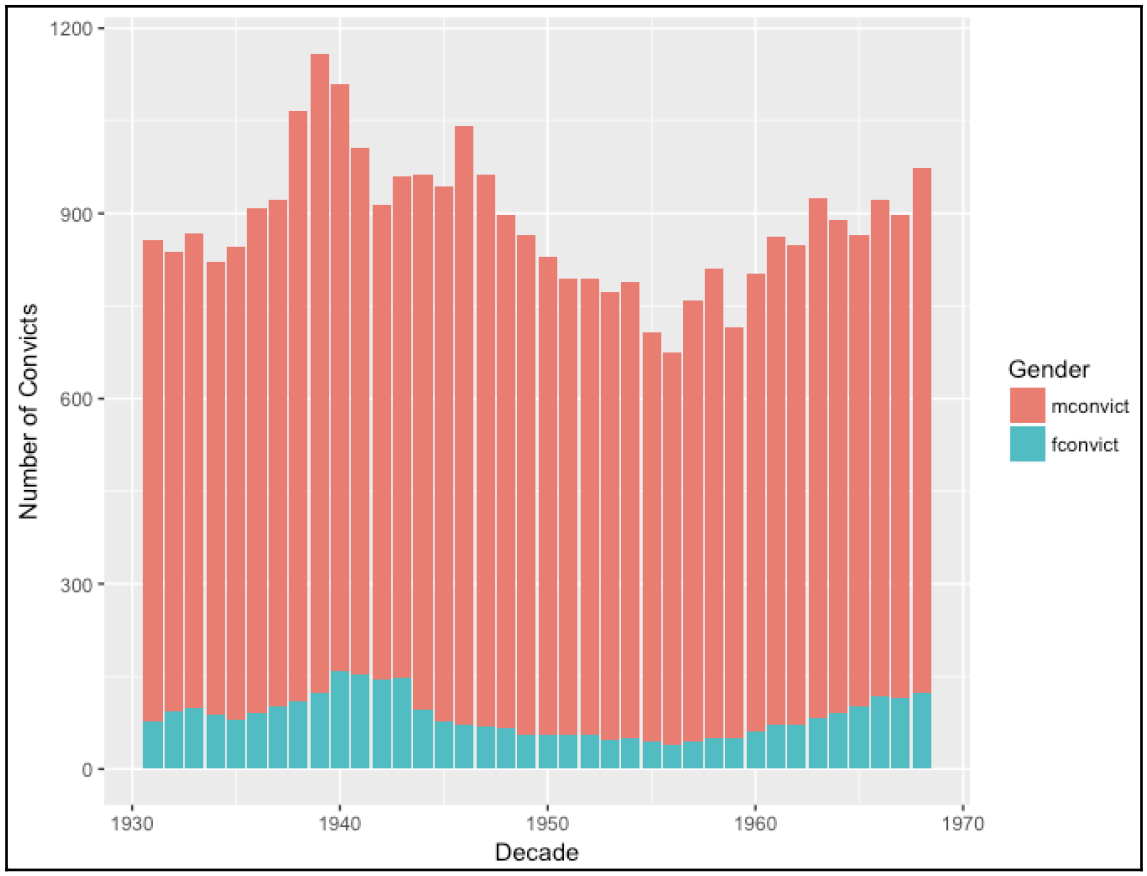

Hartnagel$mtheft
180
200
220
240
260
280
300
1930
1940
1950
1960
Hartnagel$year

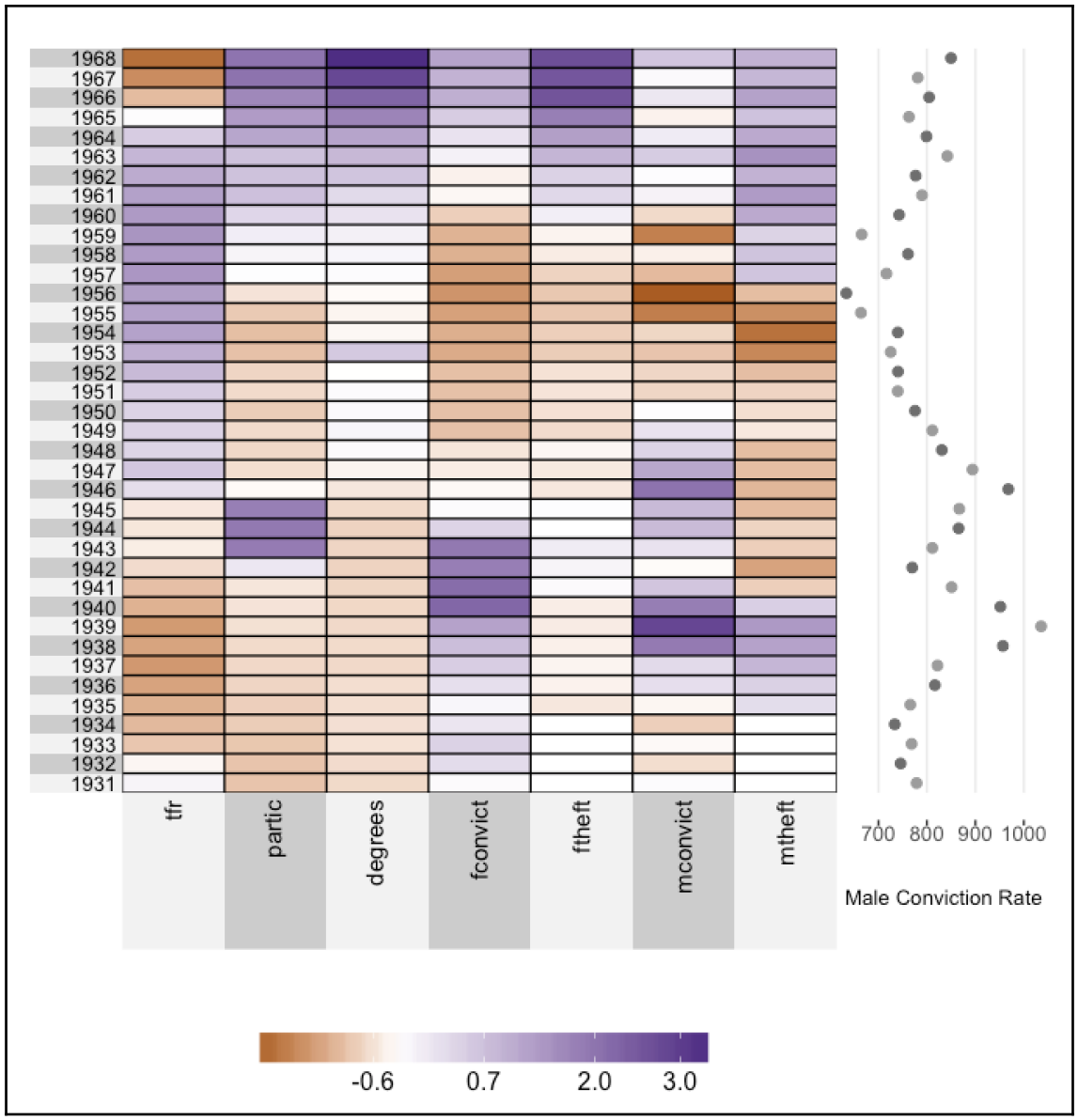

Hartnagel$mtheft
180
200
220
240
260
280
300
1930
1940
1950
1960
Hartnagel$year

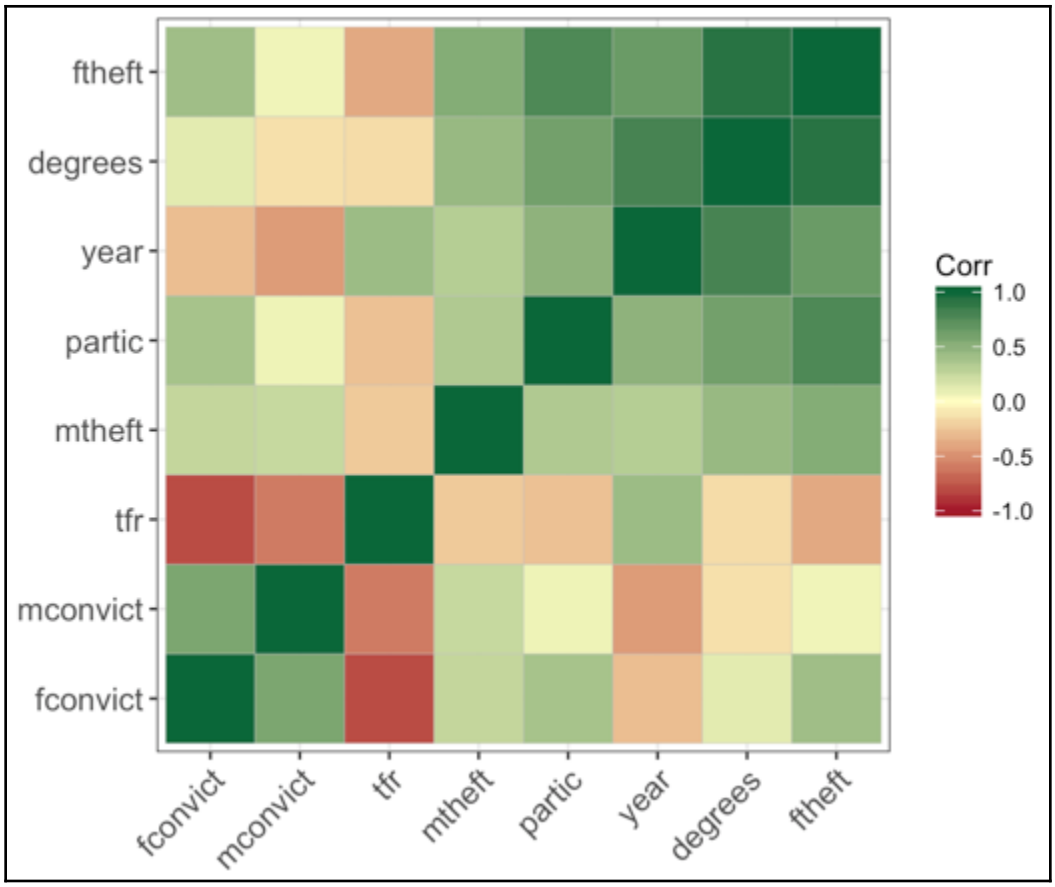

100000
150000
200000
AsstProf
AssocProf
Prof

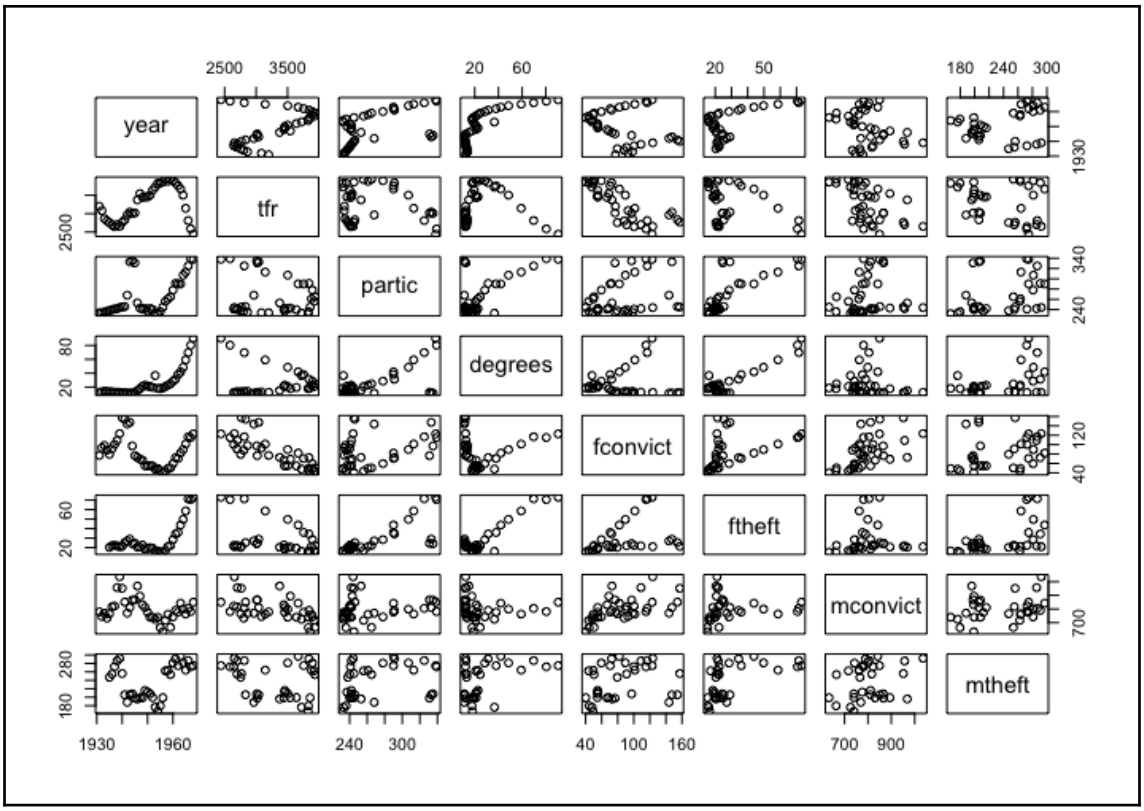

1000
750
500
250
0
Number of Convicts
1930
1940
1950
1960
1970
Decade
Decade
1930
1940
1950
1960

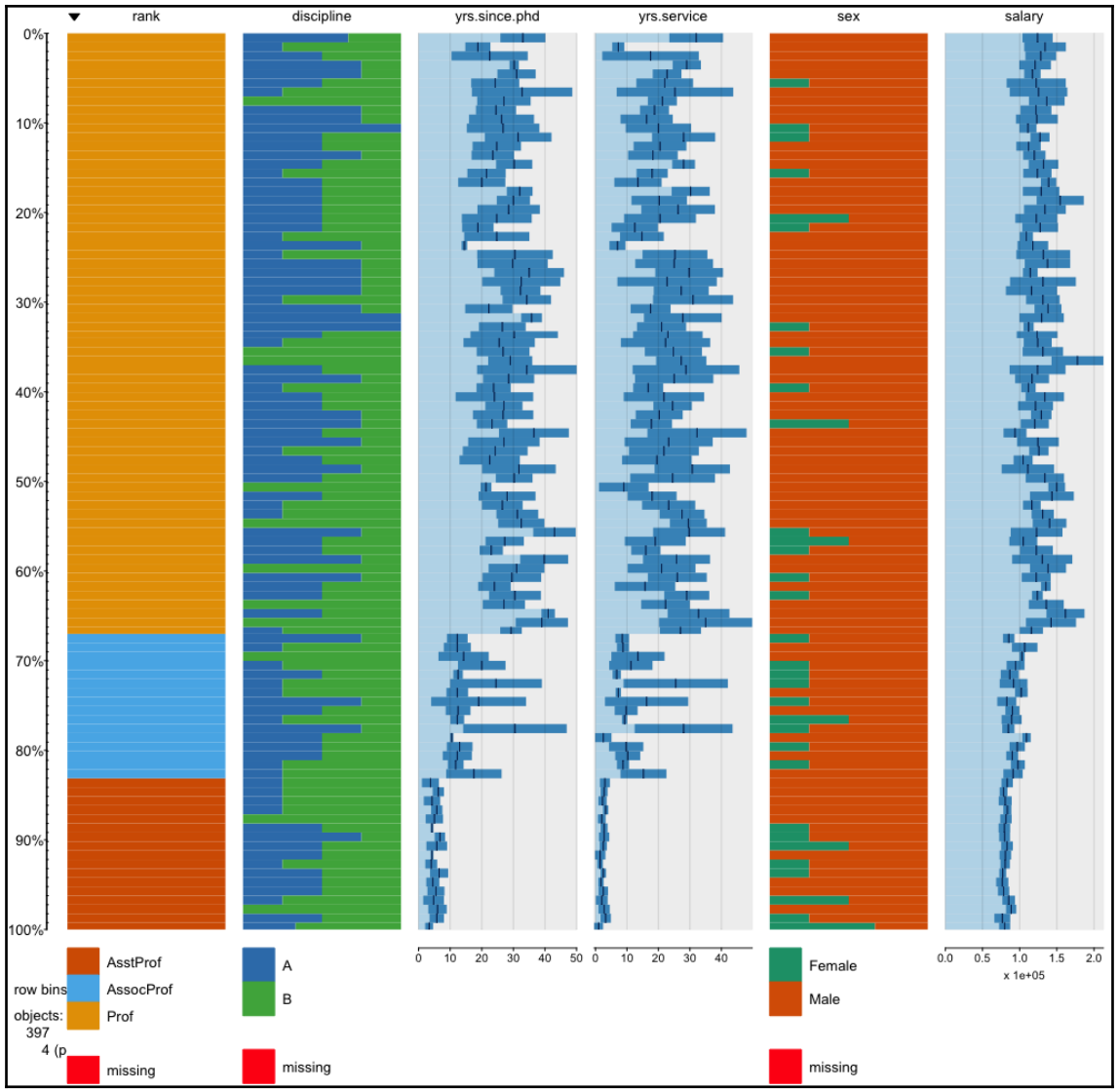

1970
1960
1950
1940
1930
Decade
0
250
500
750
1000
Male indictable-offense conviction rate per 100,000
Decade
1930
1940
1950
1960


1200
900
600
300
0
Number of Convicts
1930
1940
1950
1960
1970
Decade
Gender
mconvict
fconvict

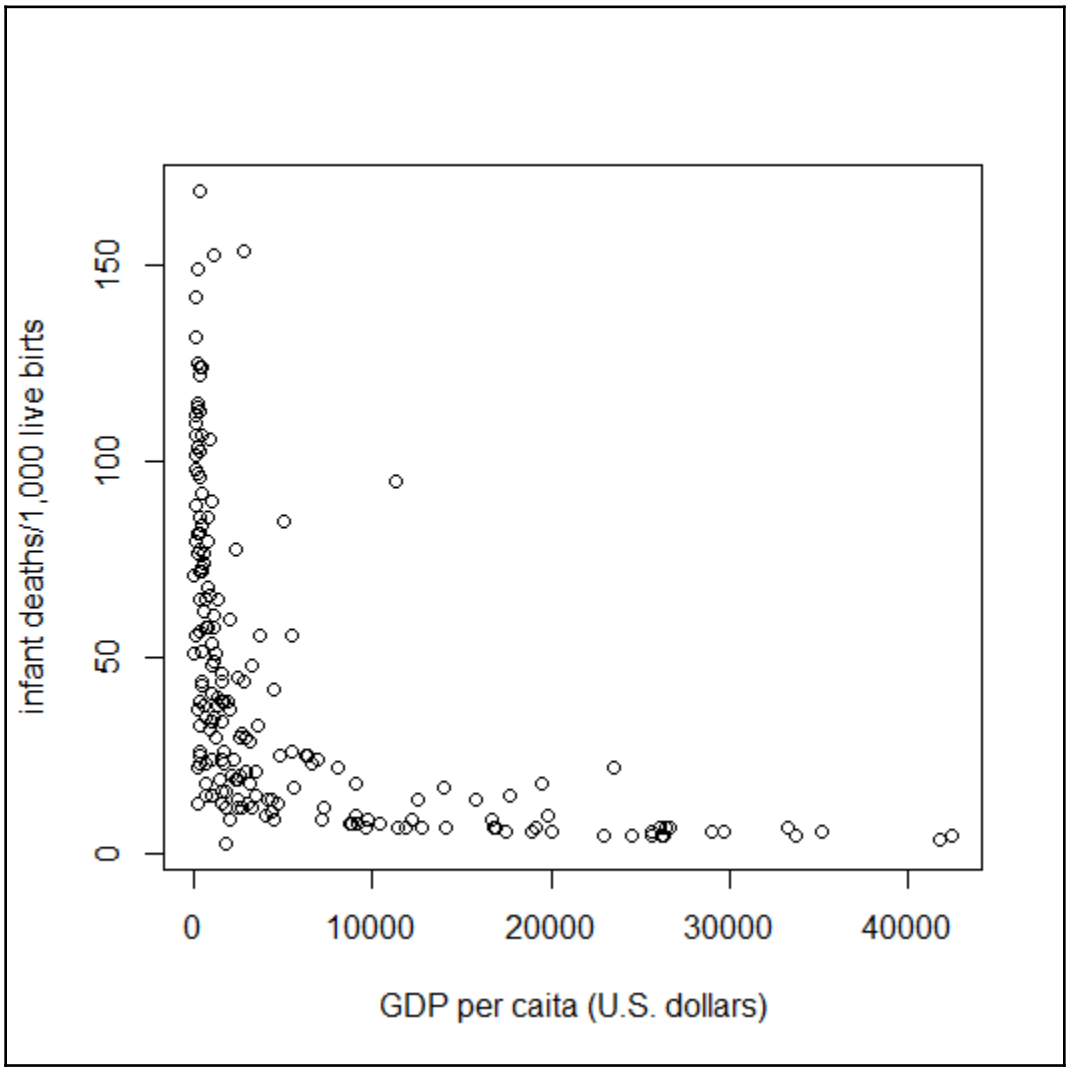

1968
1967
1966
1965
1964
1963
1962
1961
1960
1959
1958
1957
1956
1955
1954
1953
1952
1951
1950
1949
1948
1947
1946
1945
1944
1943
1942
1941
1940
1939
1938
1937
1936
1935
1934
1933
1932
1931
tfr
partic
degrees
fconvict
ftheft
mconvict
mtheft
700
800
900
1000
Male Conviction Rate
-0.6
0.7
2.0
3.0

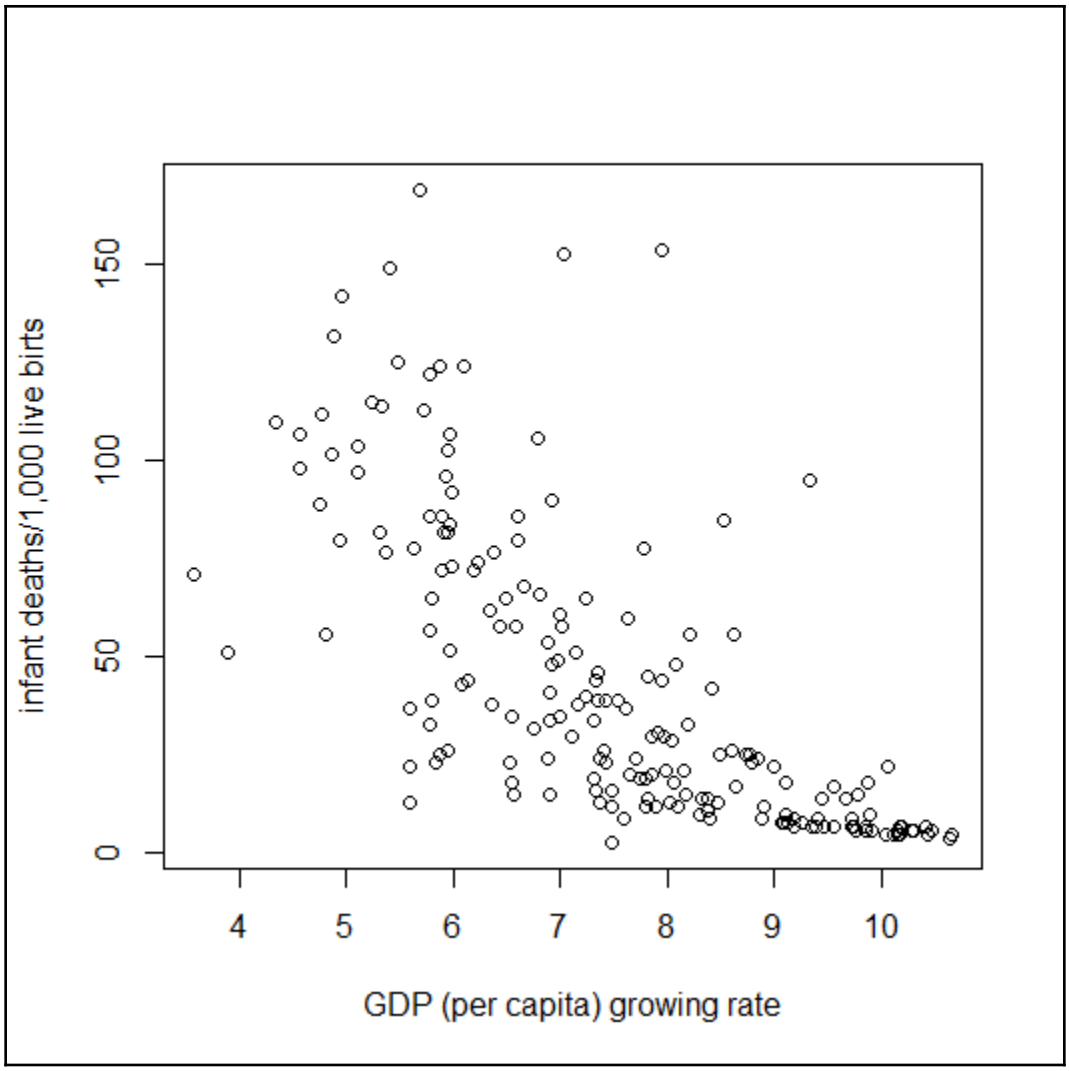

ftheft
degrees
year
partic
mtheft
tfr
mconvict
fconvict
fconvict
mconvict
tfr
mtheft
partic
year
degrees
ftheft
Corr
1.0
0.5
0.0
-0.5
-1.0

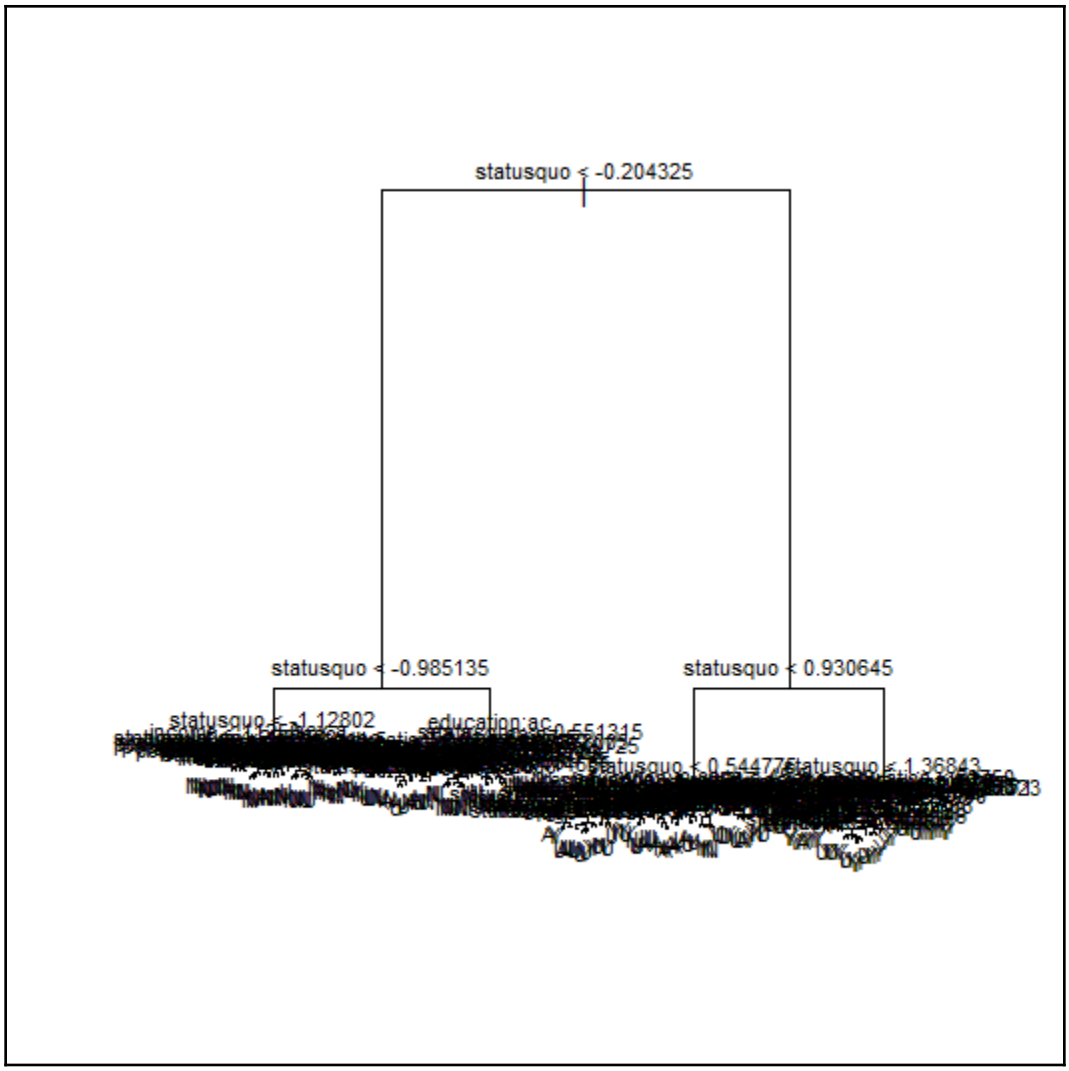

year
tfr
partic
degrees
fconvict
ftheft
mconvict
mtheft
2500 3500
20 60
20 50
180 240 300
1930
2500
340
240
80
20
120
40
60
20
700
280
180
1930 1960
240 300
40 100 160
700 900

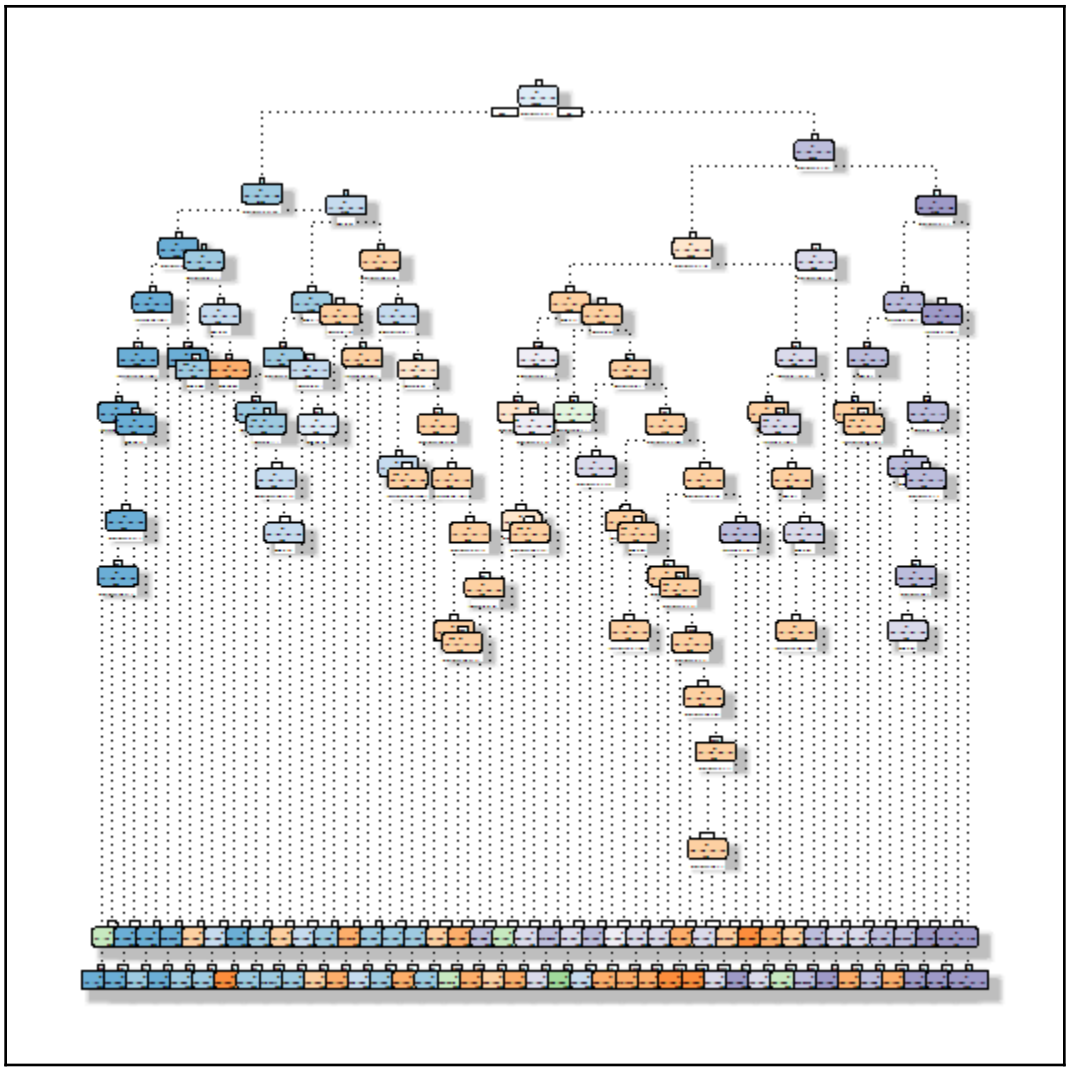

rank
discipline
yrs.since.phd
yrs.service
sex
salary
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
0 10 20 30 40 50
0 10 20 30 40
0.0 0.5 1.0 1.5 2.0
x 1e+05
row bins
objects:
397
4 (p
AsstProf
AssocProf
Prof
missing
A
B
missing
Female
Male
missing

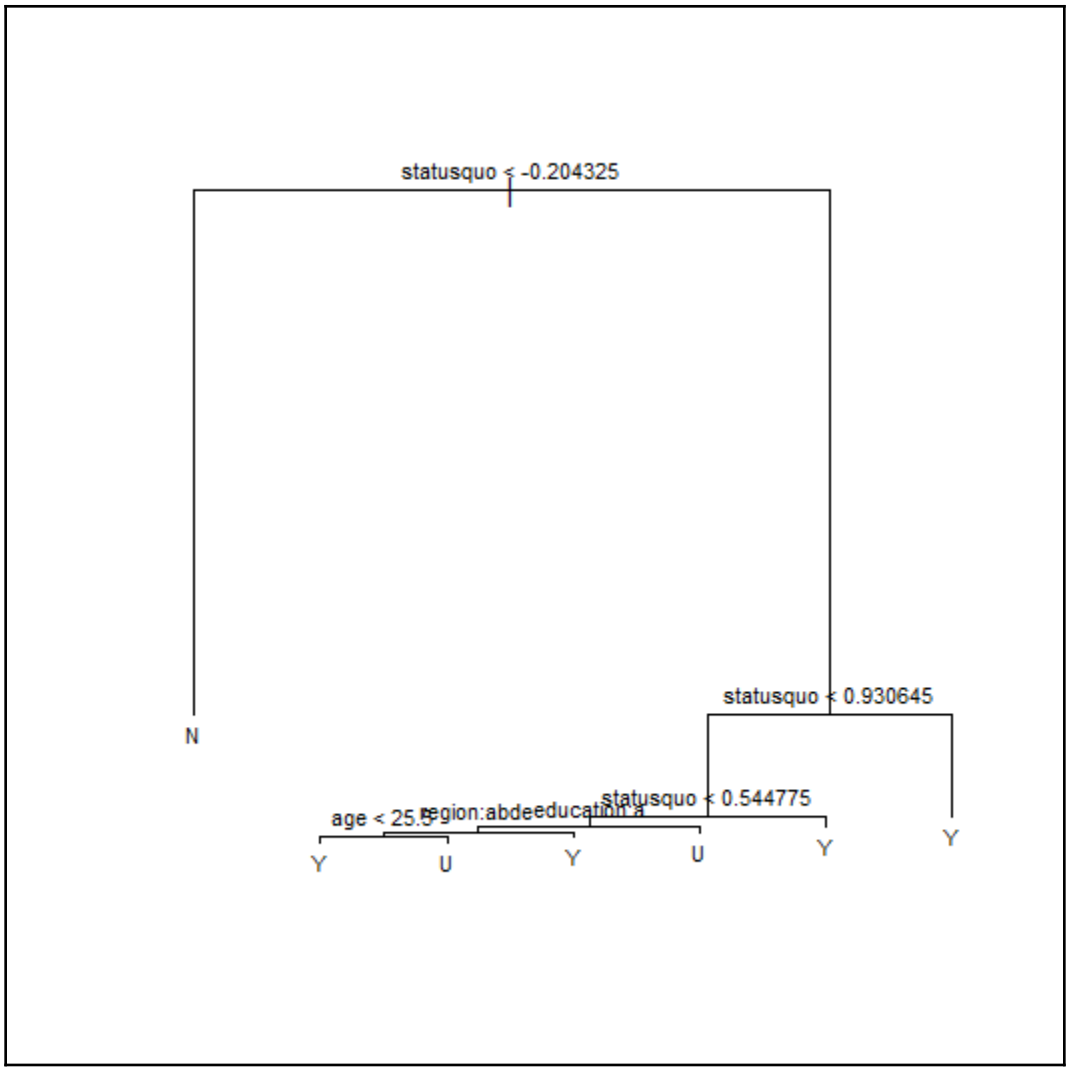
Chapter 6: Machine Learning with R

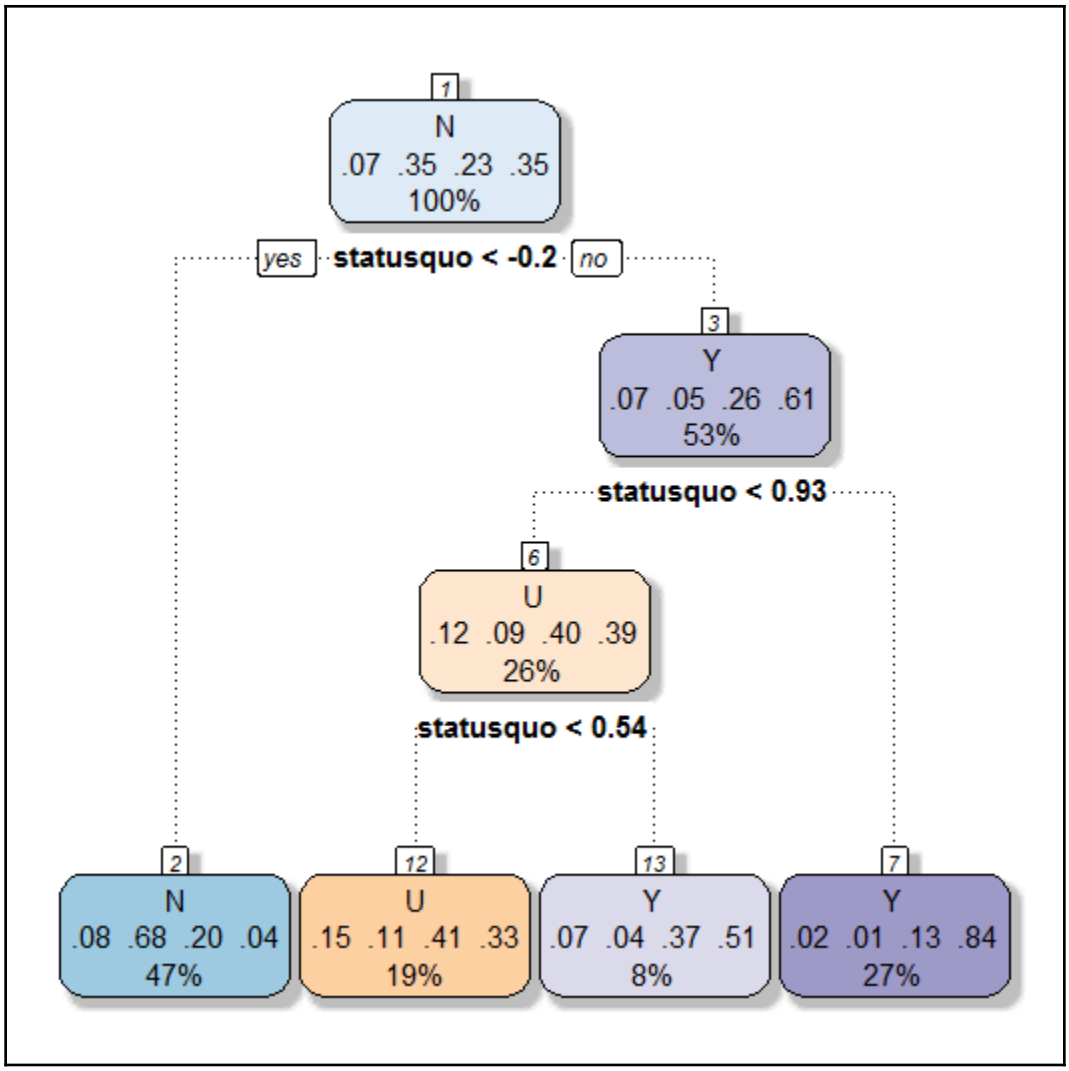



infant deaths/1,000 live birts
0
50
100
150
4
5
6
7
8
9
10
GDP (per capita) growing rate

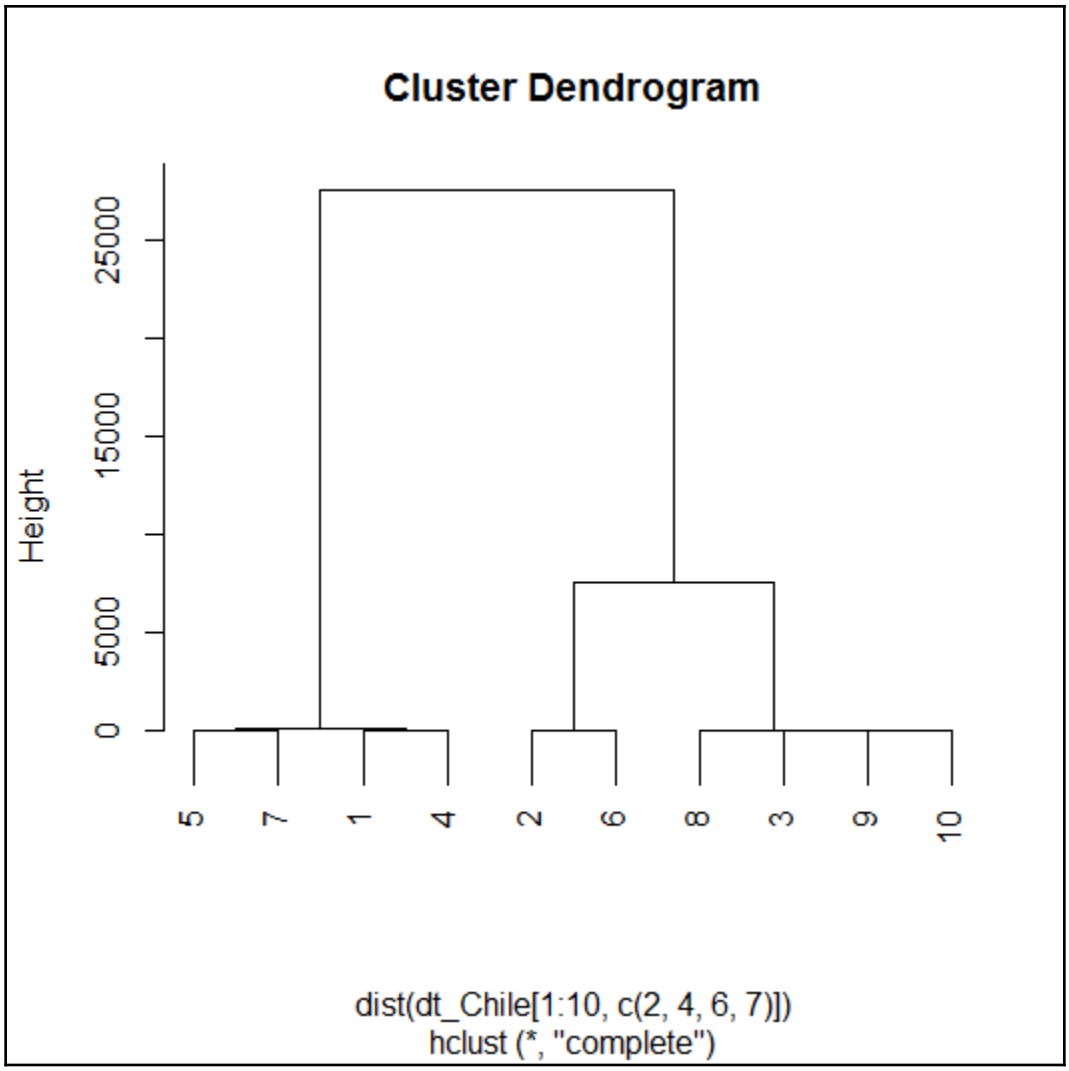

statusquo < -0.204325
statusquo < -0.985135
statusquo < 0.930645
statusquo < -1.12802
education:ac
statusquo < 0.544776
statusquo < 1.36843

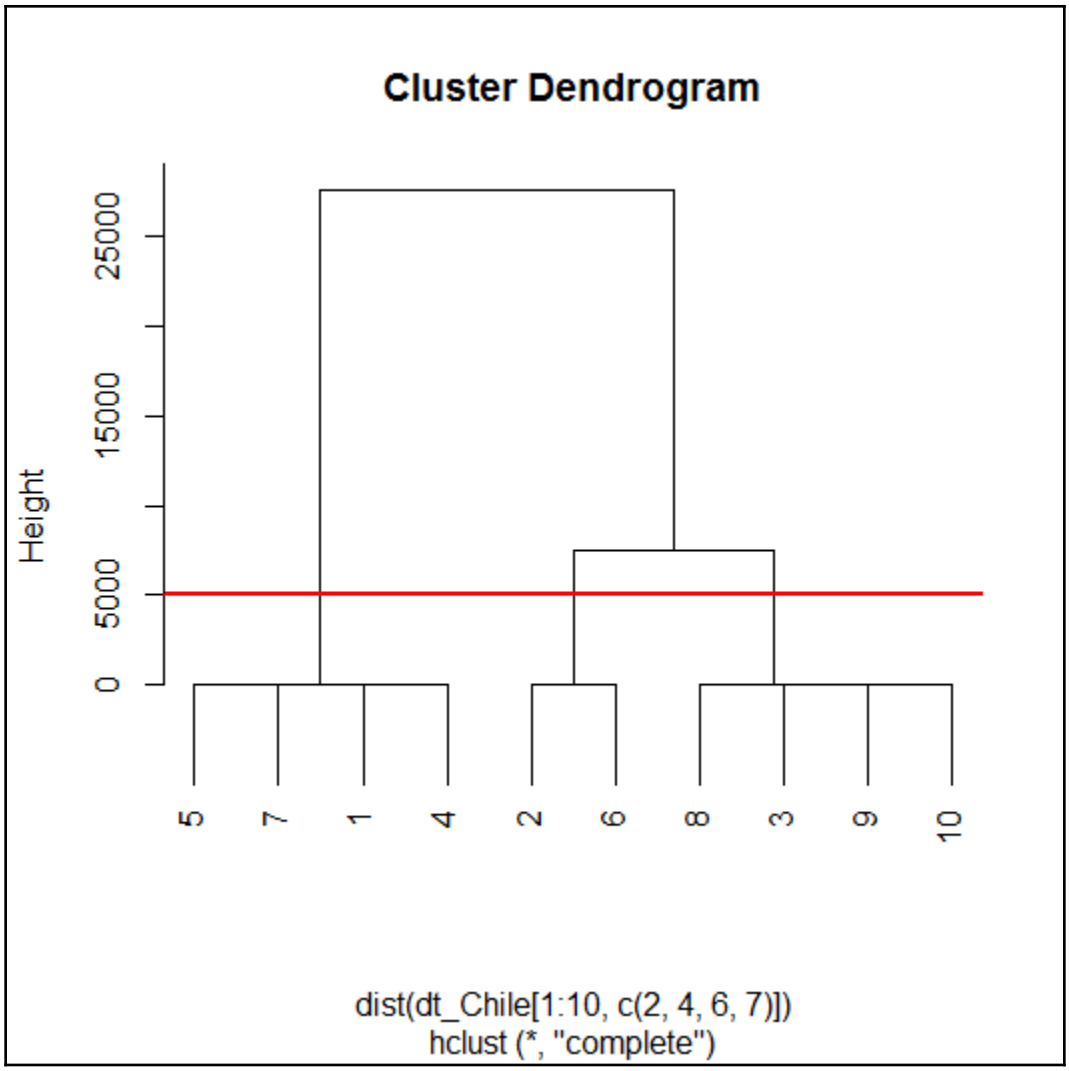


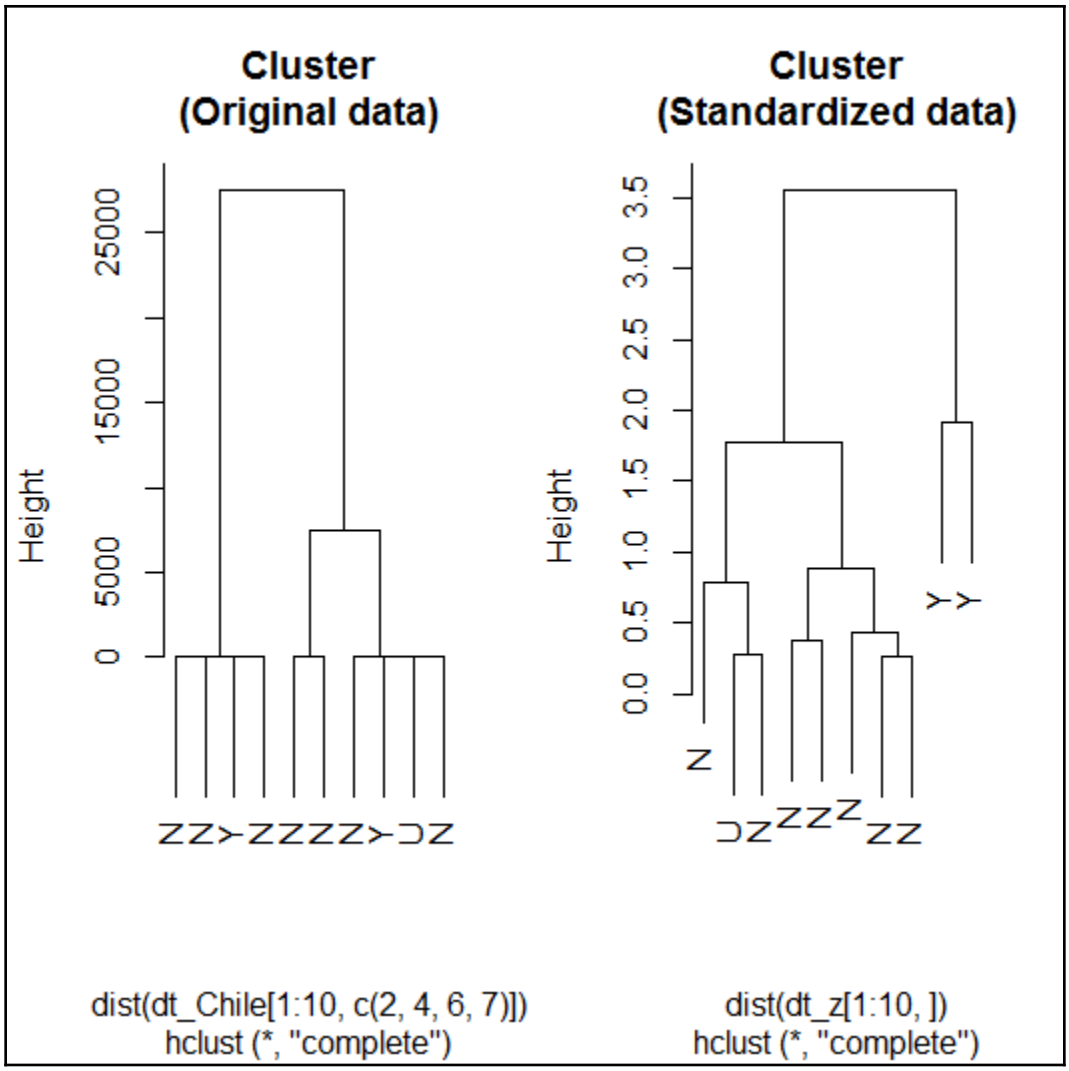

statusquo < -0.204325
N
statusquo < 0.930645
statusquo < 0.544775
age < 25.5
region:abdeeducation.a
Y
U
Y
U
Y
Y


1
N
.07 .35 .23 .35
100%
yes
statusquo < -0.2
no
3
Y
.07 .05 .26 .61
53%
statusquo < 0.93
6
U
.12 .09 .40 .39
26%
statusquo < 0.54
2
N
.08 .68 .20 .04
47%
12
U
.15 .11 .41 .33
19%
13
Y
.07 .04 .37 .51
8%
7
Y
.02 .01 .13 .84
27%

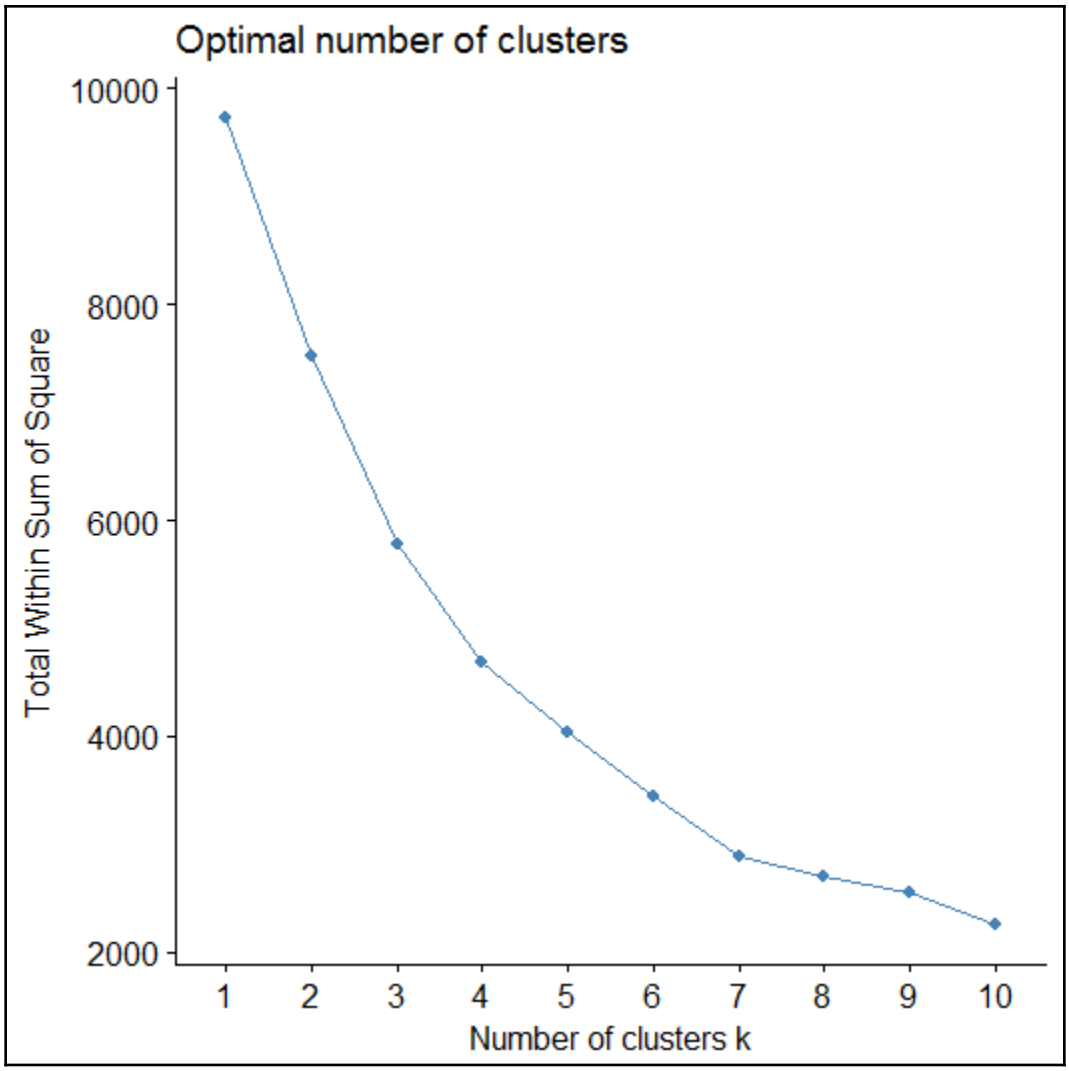
population	age	income	statusquo
Min. :175000	Min. :23.0	Min. : 7500	Min. :-1.2962
1st Qu.:175000	1st Qu.:29.0	1st Qu.:15000	1st Qu.:-1.1050
Median :175000	Median :38.0	Median :35000	Median :-1.0316
Mean :175000	Mean :40.8	Mean :25500	Mean :-0.2388
3rd Qu.:175000	3rd Qu.:49.0	3rd Qu.:35000	3rd Qu.: 1.0082
Max. :175000	Max. :65.0	Max. :35000	Max. : 1.2307


Cluster Dendrogram
Height
0
5000
15000
25000
5
7
1
4
2
6
8
3
9
10
dist(dt_Chile[1:10, c(2, 4, 6, 7)])
hclust (*, "complete")

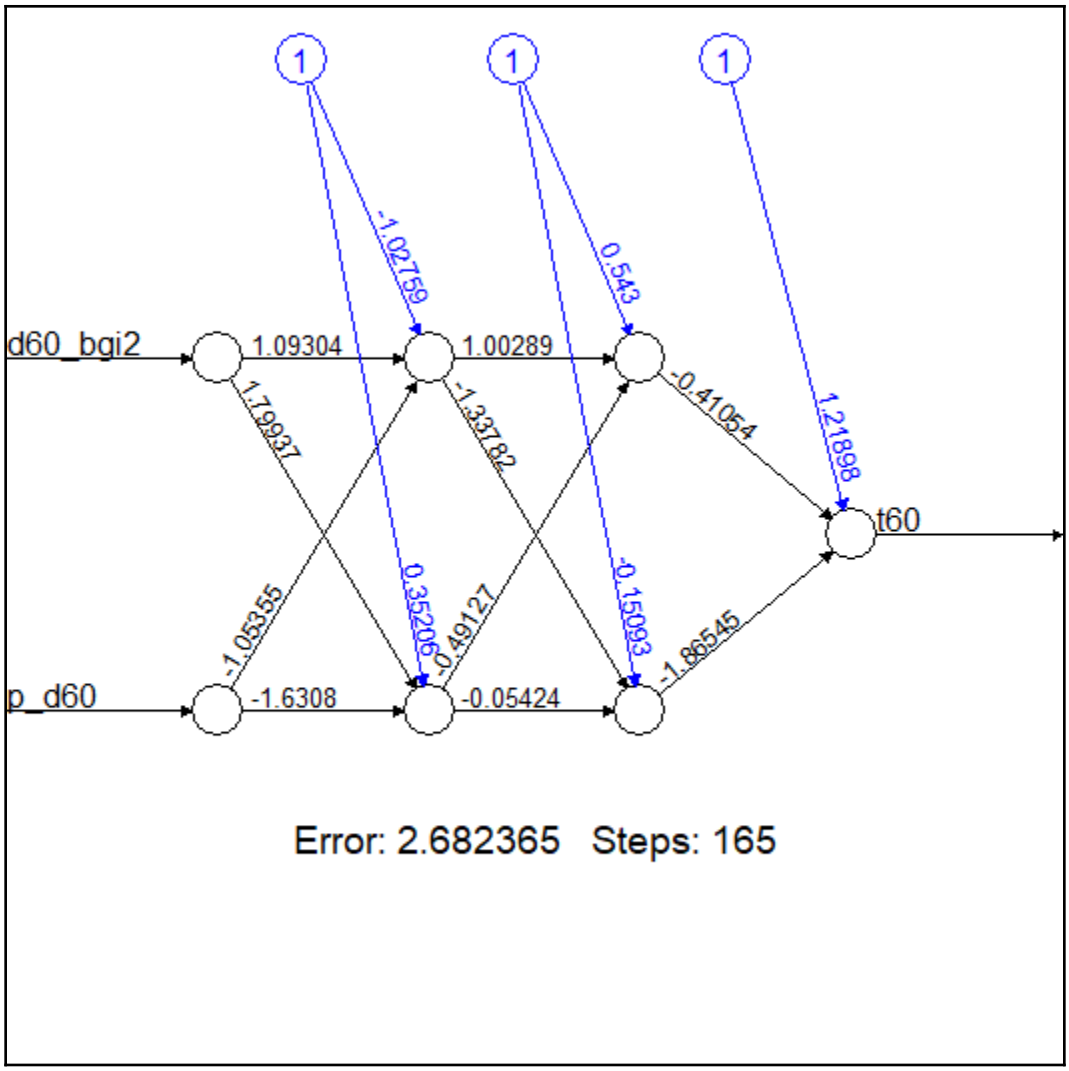

Cluster Dendrogram
Height
0
5000
15000
25000
5
7
1
4
2
6
8
3
9
10
dist(dt_Chile[1:10, c(2, 4, 6, 7)])
hclust (*, "complete")


Cluster
(Original data)
Height
0
5000
15000
25000
dist(dt_Chile[1:10, c(2, 4, 6, 7)])
hclust (*, "complete")
Cluster
(Standardized data)
Height
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
dist(dt_z[1:10,])
hclust (*, "complete")

	A	N	U	Y
1	13	59	114	194
2	49	300	123	66
3	41	120	116	213
4	24	112	43	115

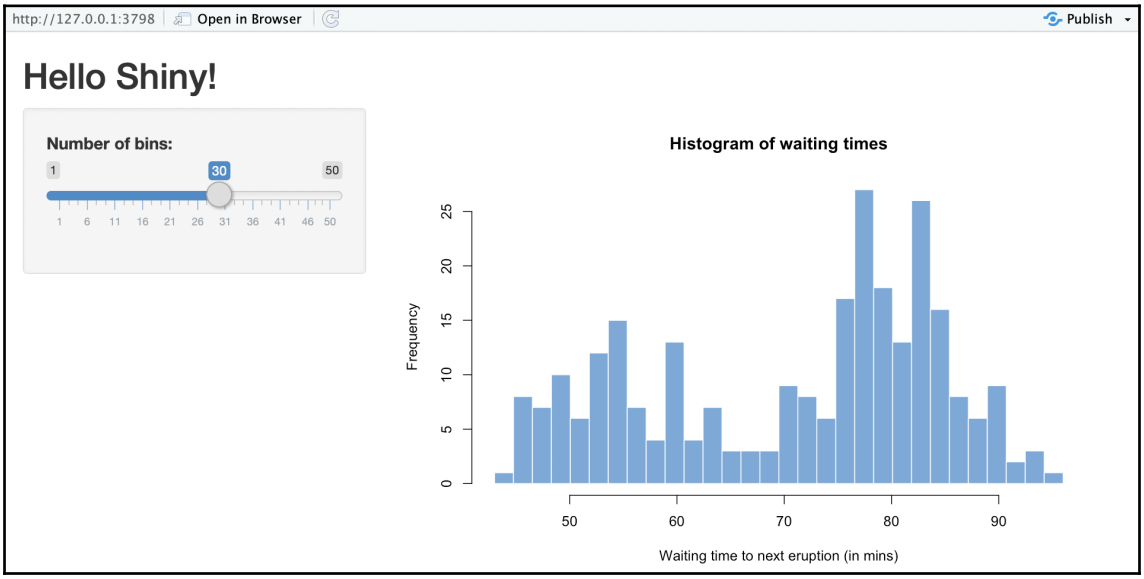


clusters	region	population	sex	age	education	income	statusquo	vote	hc_cluster
1	SA	189620.23	M	36.19253	S	25402.11	-0.7710496	N	2
2	S	87112.64	F	40.58512	P	22178.90	0.8636202	Y	3
3	SA	226815.07	F	39.79909	PS	135502.28	0.3404584	Y	4


1
1
1
-1.02759
0.543
d60_bgi2
1.09304
1.00289
-0.41054
1.21898
1.79937
-1.33782
t60
-1.05355
0.35206
-0.49127
-0.15093
-1.86545
p_d60
-1.6308
-0.05424
Error: 2.682365 Steps: 165


Scoring History
Training
Validation
classification_error
0.26
0.28
0.30
0.32
0.34
0.36
0
10
20
30
40
epochs

Chapter 7: Forecasting and ML App with R



data.gov.uk | Find open data

Publish your data Support

We've been improving data.gov.uk to help you find and use open government data.
Discover what's changed and get in touch to give us your feedback.

Don't show this message again

Home > NHS Digital > GP practice prescribing data - Presentation level

GP practice prescribing data - Presentation level

Published by: NHS Digital
Last updated: 15 October 2018
Topic: Health
Licence: Open Government Licence

Summary

Warning: Large file size (over 1GB).

Each monthly data set is large (over 4 million rows), but can be viewed in standard software such as Microsoft WordPad (save by right-clicking on the file name and selecting 'Save Target As', or equivalent on Mac OSX). It is then possible to select the required rows of data and copy and paste the information into another

View full summary

More from this publisher
All datasets from NHS Digital

Related datasets
GP practice prescribing data - Chemical level
GP Prescribing Data
GP Practice - Demographic Data
Numbers of Patients Registered at a GP Practice

Search

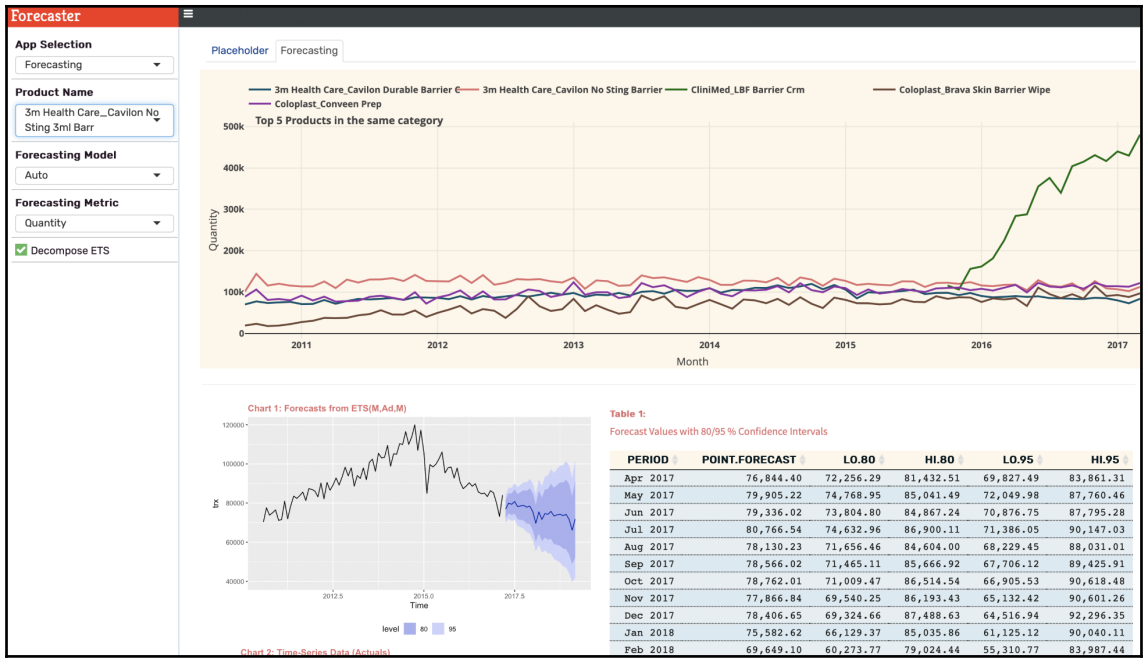
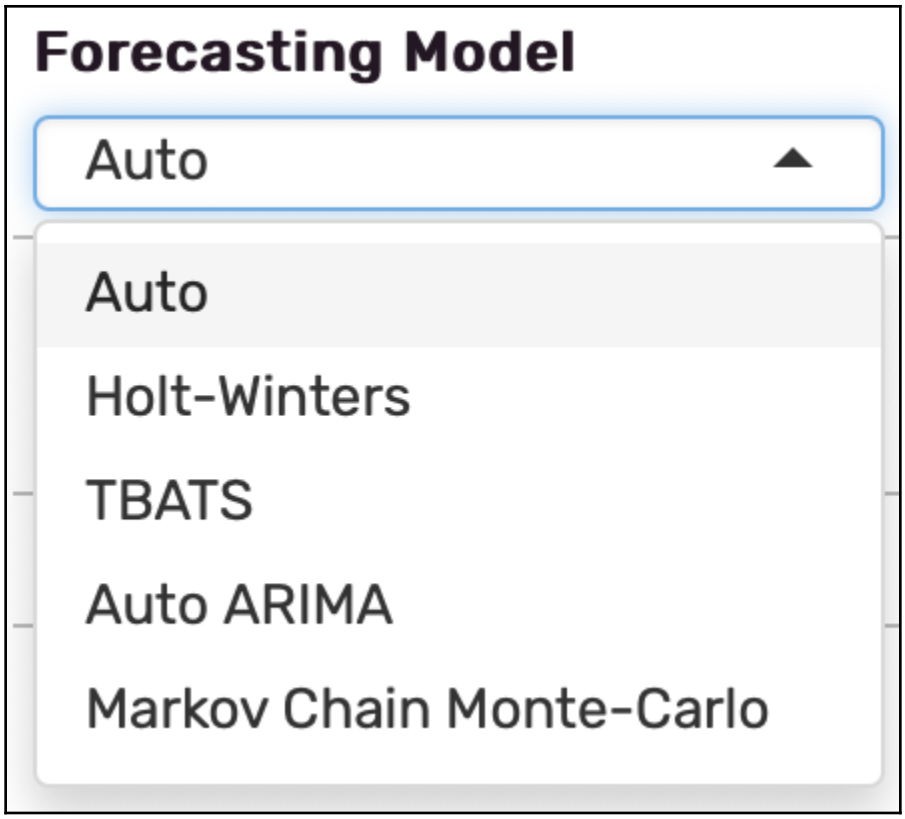


Table 1:

Forecast Values with 80/95 % Confidence Intervals

PERIOD	POINT.FORECAST	LO.80	HI.80	LO.95	HI.95
Apr 2017	76,844.40	72,256.29	81,432.51	69,827.49	83,861.31
May 2017	79,905.22	74,768.95	85,041.49	72,049.98	87,760.46
Jun 2017	79,336.02	73,804.80	84,867.24	70,876.75	87,795.28
Jul 2017	80,766.54	74,632.96	86,900.11	71,386.05	90,147.03
Aug 2017	78,130.23	71,656.46	84,604.00	68,229.45	88,031.01
Sep 2017	78,566.02	71,465.11	85,666.92	67,706.12	89,425.91
Oct 2017	78,762.01	71,009.47	86,514.54	66,905.53	90,618.48
Nov 2017	77,866.84	69,540.25	86,193.43	65,132.42	90,601.26
Dec 2017	78,406.65	69,324.66	87,488.63	64,516.94	92,296.35
Jan 2018	75,582.62	66,129.37	85,035.86	61,125.12	90,040.11
Feb 2018	69,649.10	60,273.77	79,024.44	55,310.77	83,987.44


Forecasting Model
Auto
Auto
Holt-Winters
TBATS
Auto ARIMA
Markov Chain Monte-Carlo

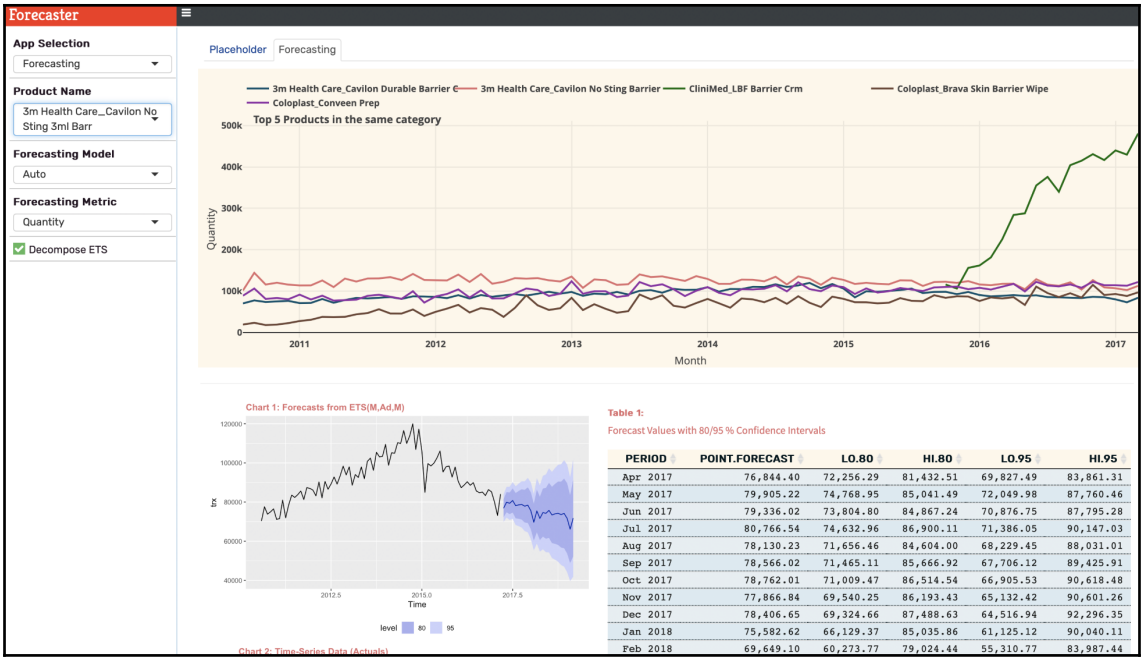


Table 1:

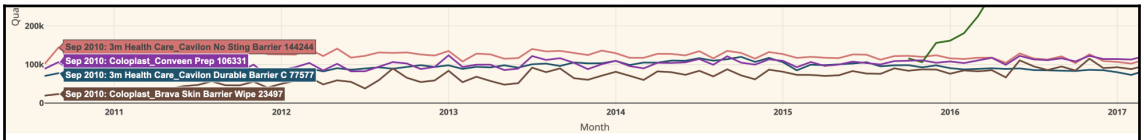
Forecast Values with 80/95 % Confidence Intervals

PERIOD	POINT.FORECAST	LO.80	HI.80	LO.95	HI.95
Apr 2017	76,844.40	72,256.29	81,432.51	69,827.49	83,861.31
May 2017	79,905.22	74,768.95	85,041.49	72,049.98	87,760.46
Jun 2017	79,336.02	73,804.80	84,867.24	70,876.75	87,795.28
Jul 2017	80,766.54	74,632.96	86,900.11	71,386.05	90,147.03
Aug 2017	78,130.23	71,656.46	84,604.00	68,229.45	88,031.01
Sep 2017	78,566.02	71,465.11	85,666.92	67,706.12	89,425.91
Oct 2017	78,762.01	71,009.47	86,514.54	66,905.53	90,618.48
Nov 2017	77,866.84	69,540.25	86,193.43	65,132.42	90,601.26
Dec 2017	78,406.65	69,324.66	87,488.63	64,516.94	92,296.35
Jan 2018	75,582.62	66,129.37	85,035.86	61,125.12	90,040.11
Feb 2018	69,649.10	60,273.77	79,024.44	55,310.77	83,987.44
Mar 2018	75,451.61	64,556.26	86,346.97	58,788.60	92,114.62
Apr 2018	71,601.47	60,544.65	82,658.29	54,691.52	88,511.42
May 2018	74,610.43	62,327.06	86,893.80	55,824.64	93,396.22
Jun 2018	74,231.34	61,239.86	87,222.81	54,362.59	94,100.09
Jul 2018	75,721.44	61,672.16	89,770.71	54,234.92	97,207.95
Aug 2018	73,393.10	58,994.12	87,792.08	51,371.77	95,414.44
Sep 2018	73,943.19	58,640.82	89,245.57	50,540.23	97,346.16
Oct 2018	74,265.38	58,090.63	90,440.13	49,528.24	99,002.52
Nov 2018	73,554.20	56,730.76	90,377.64	47,824.97	99,283.43
Dec 2018	74,194.64	56,409.46	91,979.82	46,994.56	101,394.72
Jan 2019	71,645.03	53,680.12	89,609.94	44,170.08	99,119.98
Feb 2019	66,130.87	48,815.96	83,445.79	39,650.00	92,611.75
Mar 2019	71,756.67	52,171.59	91,341.74	41,803.88	101,709.46

Table 2:

Sales Data from NHS Records

MONTH	BNFNAME	ACTCOST	QUANTITY	CHEMSUB	METRIC
2010-08	3m Health Care_Cavilon No Sting 3ml Barr	54518.58	7559	Skin Fillers And Protectives	7559
2010-09	3m Health Care_Cavilon No Sting 3ml Barr	61434.85	8527	Skin Fillers And Protectives	8527
2010-10	3m Health Care_Cavilon No Sting 3ml Barr	52582.01	7292	Skin Fillers And Protectives	7292
2010-11	3m Health Care_Cavilon No Sting 3ml Barr	54820.75	7599	Skin Fillers And Protectives	7599
2010-12	3m Health Care_Cavilon No Sting 3ml Barr	50962.04	7061	Skin Fillers And Protectives	7061
2011-01	3m Health Care_Cavilon No Sting 3ml Barr	53345.9	7394	Skin Fillers And Protectives	7394
2011-02	3m Health Care_Cavilon No Sting 3ml Barr	49534.52	6858	Skin Fillers And Protectives	6858
2011-03	3m Health Care_Cavilon No Sting 3ml Barr	55615.69	7700	Skin Fillers And Protectives	7700
2011-04	3m Health Care_Cavilon No Sting 3ml Barr	49971.22	6947	Skin Fillers And Protectives	6947
2011-05	3m Health Care_Cavilon No Sting 3ml Barr	54943.66	7604	Skin Fillers And Protectives	7604
2011-06	3m Health Care_Cavilon No Sting 3ml Barr	55323.1	7661	Skin Fillers And Protectives	7661
2011-07	3m Health Care_Cavilon No Sting 3ml Barr	54619.06	7565	Skin Fillers And Protectives	7565
2011-08	3m Health Care_Cavilon No Sting 3ml Barr	53093.28	7352	Skin Fillers And Protectives	7352
2011-09	3m Health Care_Cavilon No Sting 3ml Barr	54413.82	7535	Skin Fillers And Protectives	7535
2011-10	3m Health Care_Cavilon No Sting 3ml Barr	48704.49	6746	Skin Fillers And Protectives	6746
2011-11	3m Health Care_Cavilon No Sting 3ml Barr	52585.04	7281	Skin Fillers And Protectives	7281
2011-12	3m Health Care_Cavilon No Sting 3ml Barr	52632.39	7276	Skin Fillers And Protectives	7276
2012-01	3m Health Care_Cavilon No Sting 3ml Barr	49459.28	6861	Skin Fillers And Protectives	6861
2012-02	3m Health Care_Cavilon No Sting 3ml Barr	47578.07	6576	Skin Fillers And Protectives	6576
2012-03	3m Health Care_Cavilon No Sting 3ml Barr	49792.63	6881	Skin Fillers And Protectives	6881
2012-04	3m Health Care_Cavilon No Sting 3ml Barr	44710.94	6191	Skin Fillers And Protectives	6191
2012-05	3m Health Care_Cavilon No Sting 3ml Barr	52084.8	7199	Skin Fillers And Protectives	7199
2012-06	3m Health Care_Cavilon No Sting 3ml Barr	47665.67	6602	Skin Fillers And Protectives	6602
2012-07	3m Health Care_Cavilon No Sting 3ml Barr	49825.06	6890	Skin Fillers And Protectives	6890
2012-08	3m Health Care_Cavilon No Sting 3ml Barr	48028.42	6642	Skin Fillers And Protectives	6642
2012-09	3m Health Care_Cavilon No Sting 3ml Barr	45282.66	6273	Skin Fillers And Protectives	6273
2012-10	3m Health Care_Cavilon No Sting 3ml Barr	46453.93	6431	Skin Fillers And Protectives	6431
2012-11	3m Health Care_Cavilon No Sting 3ml Barr	47591.56	6593	Skin Fillers And Protectives	6593
2012-12	3m Health Care_Cavilon No Sting 3ml Barr	46444.13	6431	Skin Fillers And Protectives	6431
2013-01	3m Health Care_Cavilon No Sting 3ml Barr	46817.73	6473	Skin Fillers And Protectives	6473



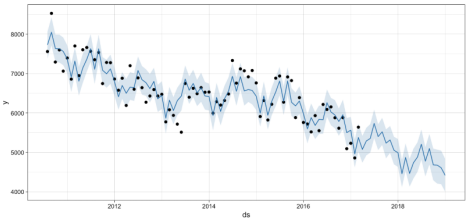
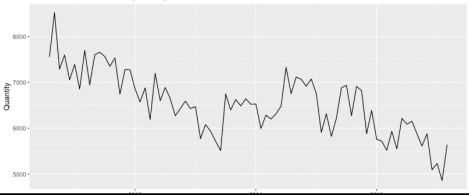


Chart 2: Time-Series Data (Actuals)

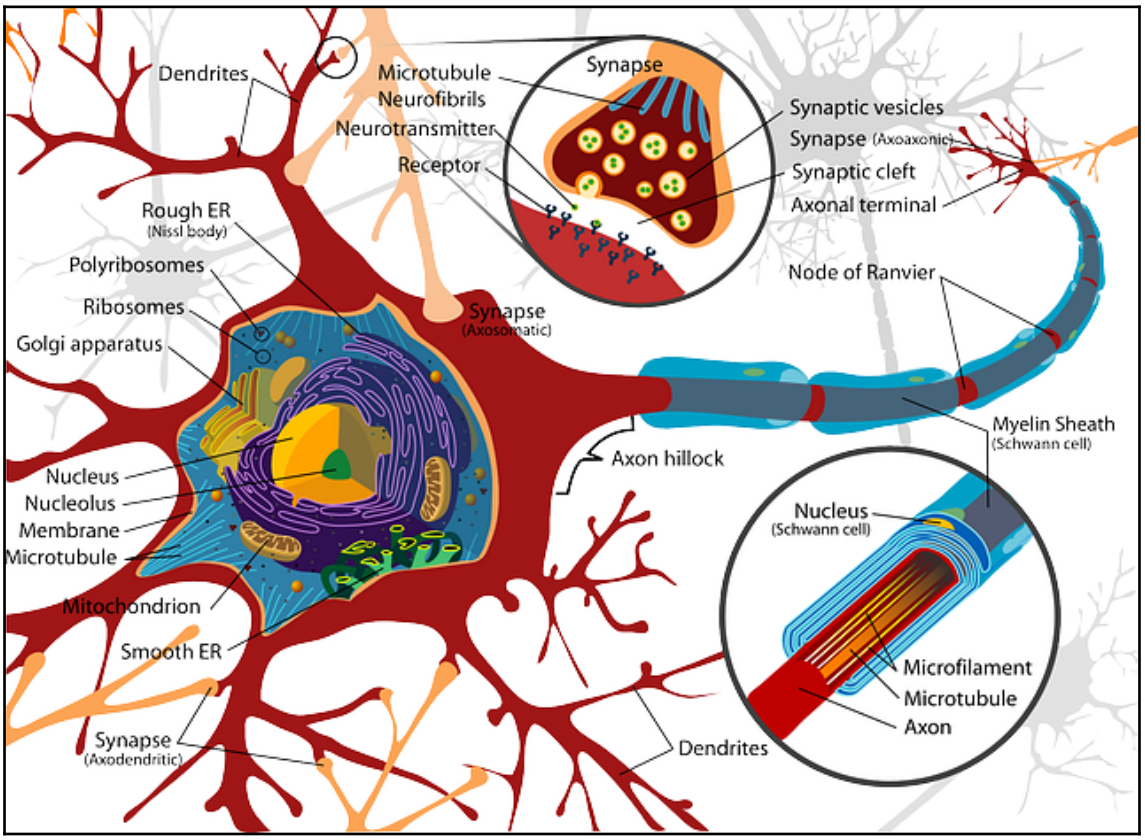


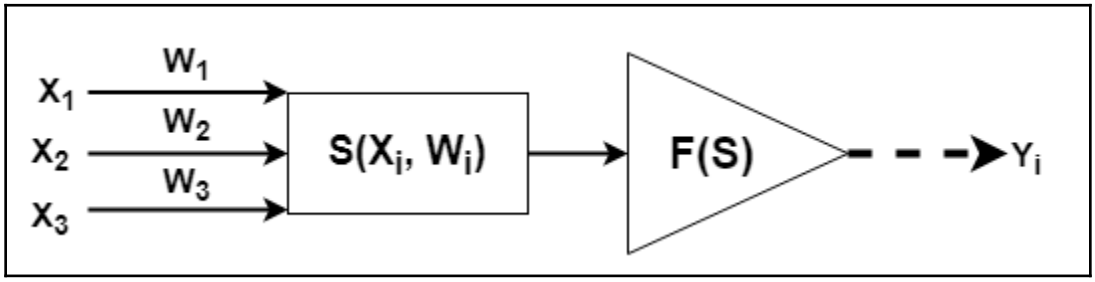
ble 1:

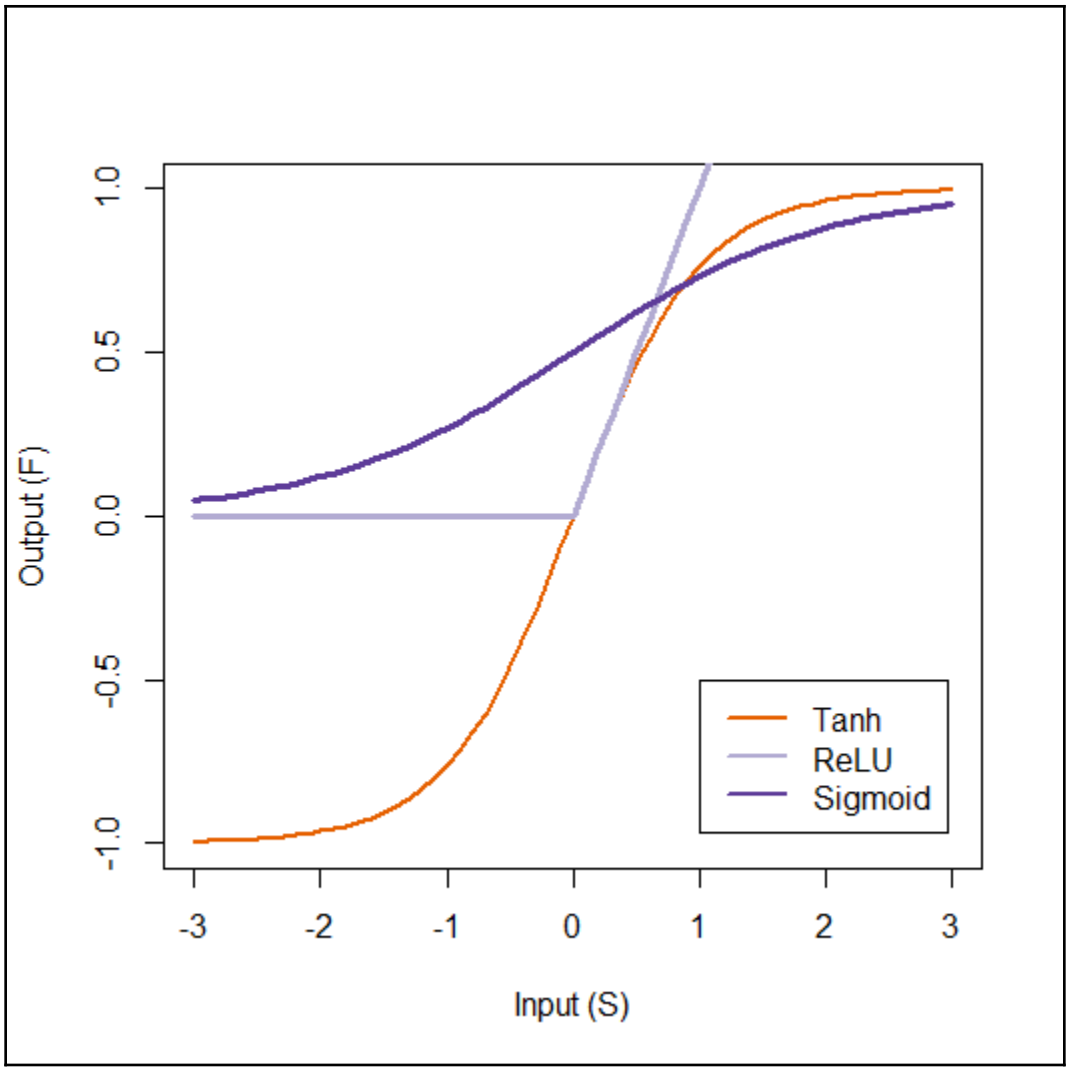
recast Values with 80/95 % Confidence Intervals

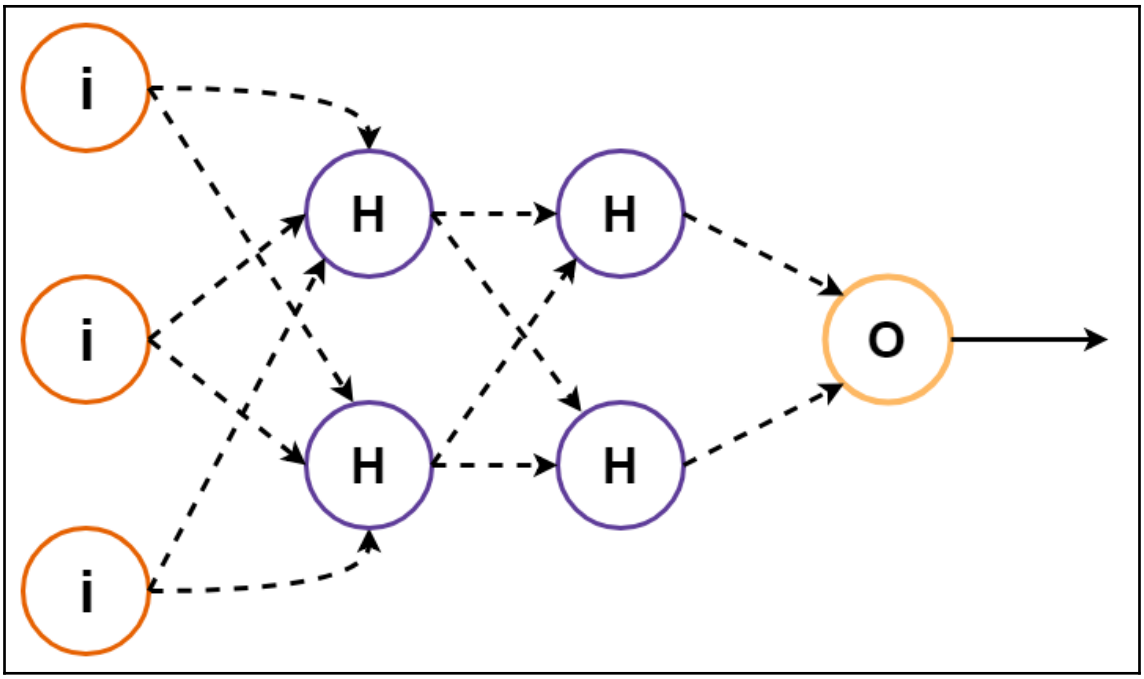
ERIOD	FCST	FCST_LOWER	FCST_UPPER	TREND	SEASONAL	SEASONAL_LOWER	SEASONAL_UPPER
2017-03	5,378.39	5,020.61	5,730.09	5,487.73	-109.34	-109.34	-109.34
2017-04	5,075.39	4,739.07	5,416.44	5,442.31	-366.92	-366.92	-366.92
2017-05	5,288.08	4,922.80	5,621.19	5,398.35	-110.27	-110.27	-110.27
2017-06	5,354.72	4,994.38	5,718.44	5,352.93	1.80	1.80	1.80
2017-07	5,734.14	5,361.38	6,083.10	5,308.97	425.17	425.17	425.17
2017-08	5,401.89	5,060.35	5,713.90	5,263.55	138.34	138.34	138.34
2017-09	5,521.54	5,186.79	5,876.15	5,218.12	303.42	303.42	303.42
2017-10	5,235.78	4,851.06	5,585.16	5,174.17	61.62	61.62	61.62
2017-11	5,315.50	4,987.56	5,668.16	5,128.74	186.76	186.76	186.76
2017-12	5,055.57	4,697.60	5,396.48	5,084.79	-29.22	-29.22	-29.22
2018-01	4,990.43	4,631.72	5,338.10	5,039.37	-48.93	-48.93	-48.93
2018-02	4,459.67	4,117.77	4,807.14	4,993.94	-534.28	-534.28	-534.28
2018-03	4,872.27	4,532.98	5,225.64	4,952.92	-80.65	-80.65	-80.65
2018-04	4,465.22	4,131.96	4,835.45	4,907.49	-442.28	-442.28	-442.28
2018-05	4,737.19	4,340.67	5,093.85	4,863.54	-126.35	-126.35	-126.35
2018-06	4,881.38	4,506.48	5,255.50	4,818.11	63.27	63.27	63.27
2018-07	5,206.20	4,815.58	5,576.40	4,774.16	432.05	432.05	432.05
2018-08	4,778.12	4,393.52	5,132.40	4,728.74	49.39	49.39	49.39
2018-09	5,096.64	4,695.47	5,466.57	4,683.31	413.33	413.33	413.33
2018-10	4,684.18	4,293.04	5,096.17	4,639.36	44.82	44.82	44.82
2018-11	4,676.79	4,207.71	5,069.79	4,593.93	82.86	82.86	82.86
2018-12	4,609.25	4,193.46	5,044.94	4,549.98	59.27	59.27	59.27
2019-01	4,418.23	4,003.88	4,832.73	4,504.55	-86.32	-86.32	-86.32

Chapter 8: Neural Networks and Deep Learning



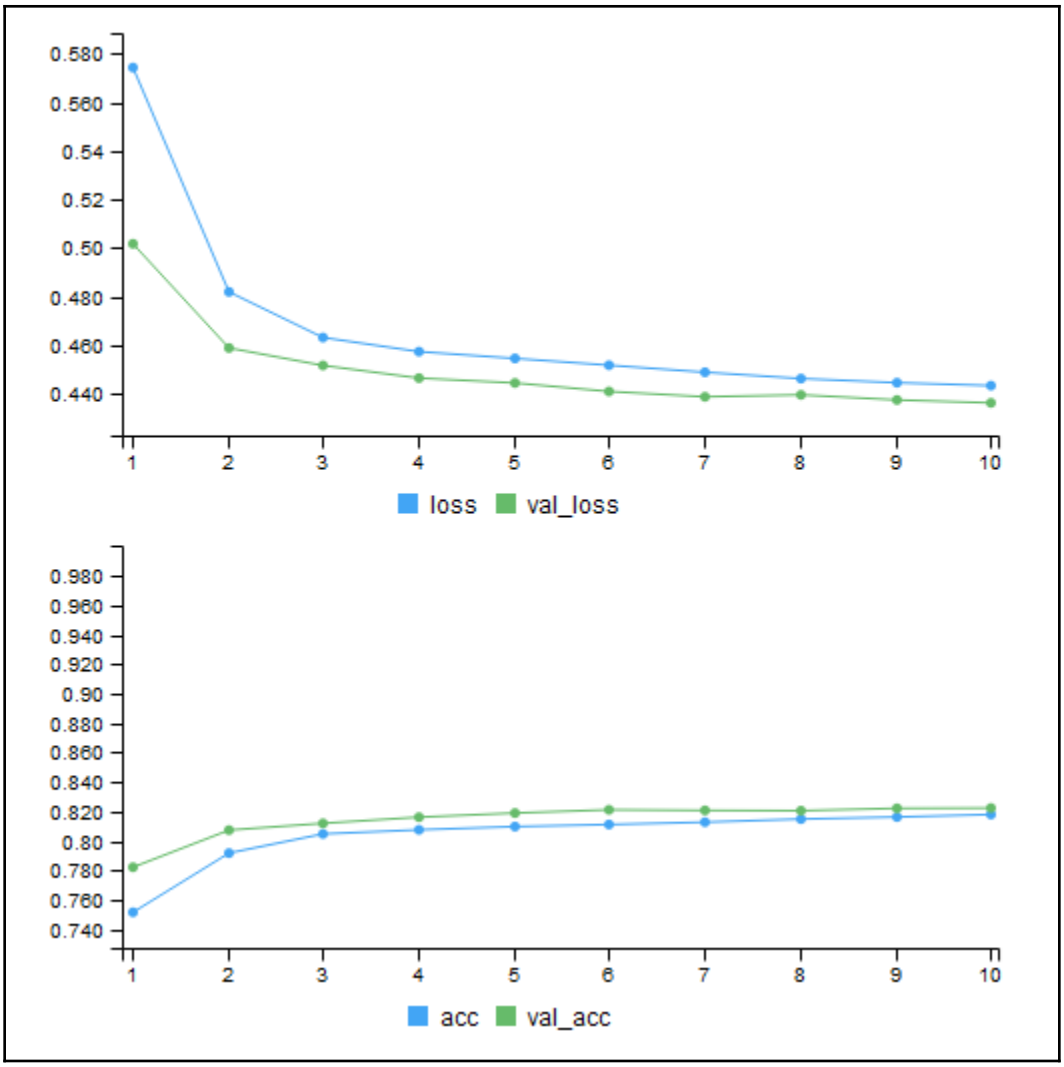



Output (F)
1.0
0.5
0.0
-0.5
-1.0
-3
-2
-1
0
1
2
3
Input (S)
Tanh
ReLU
Sigmoid

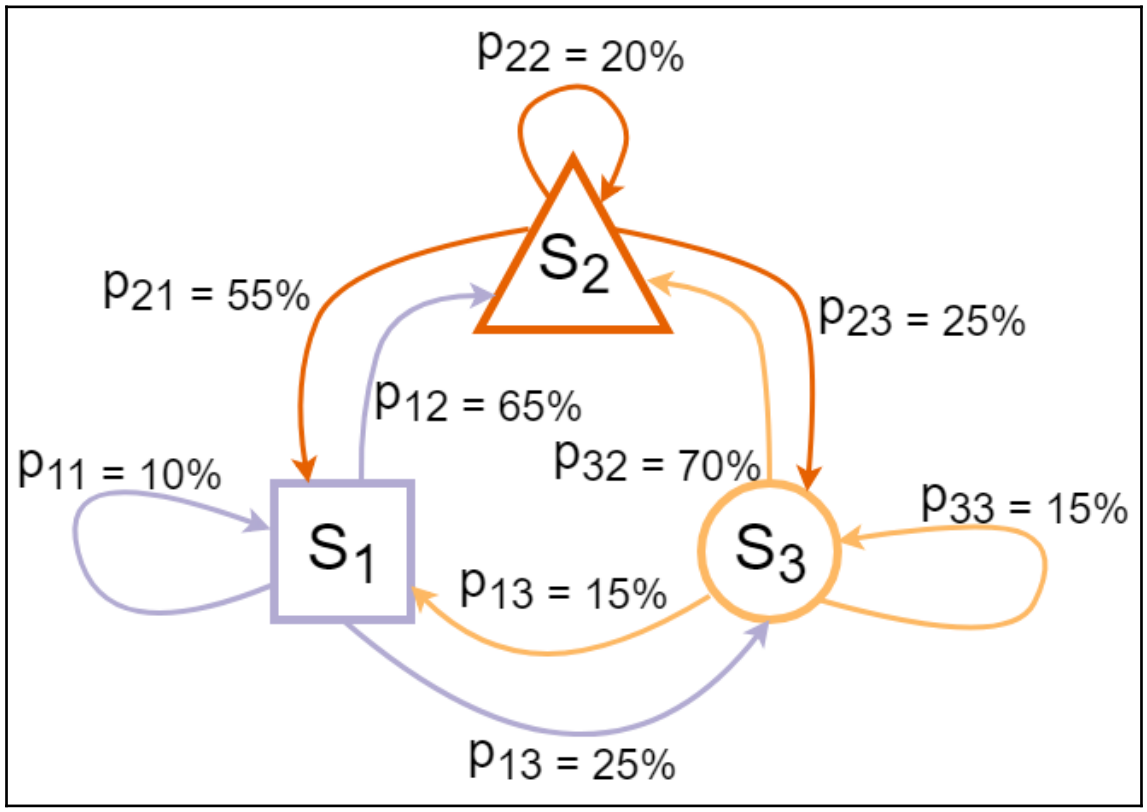


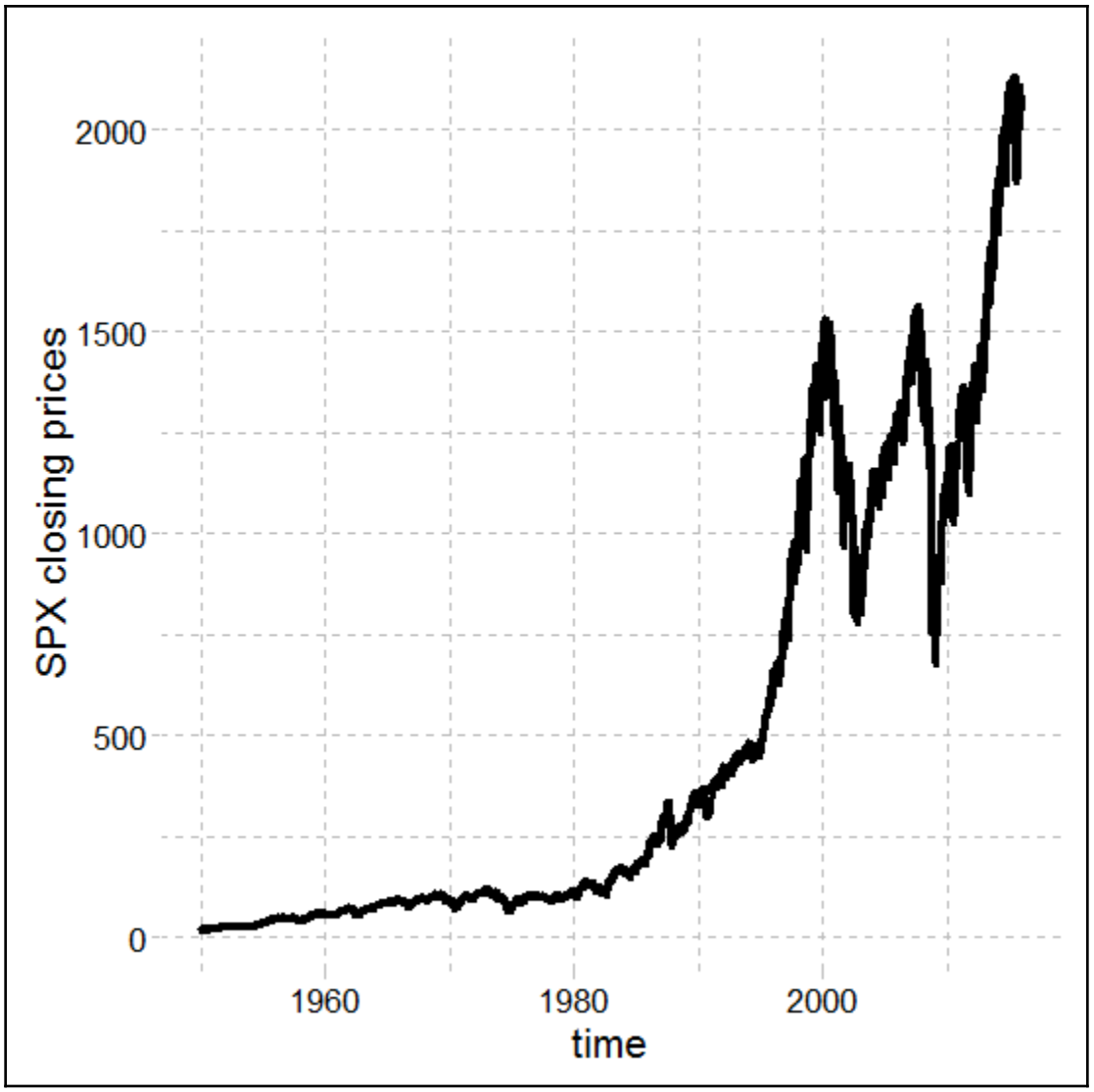
Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 25)	850
dense_2 (Dense)	(None, 15)	390
dense_3 (Dense)	(None, 6)	96
dense_4 (Dense)	(None, 2)	14

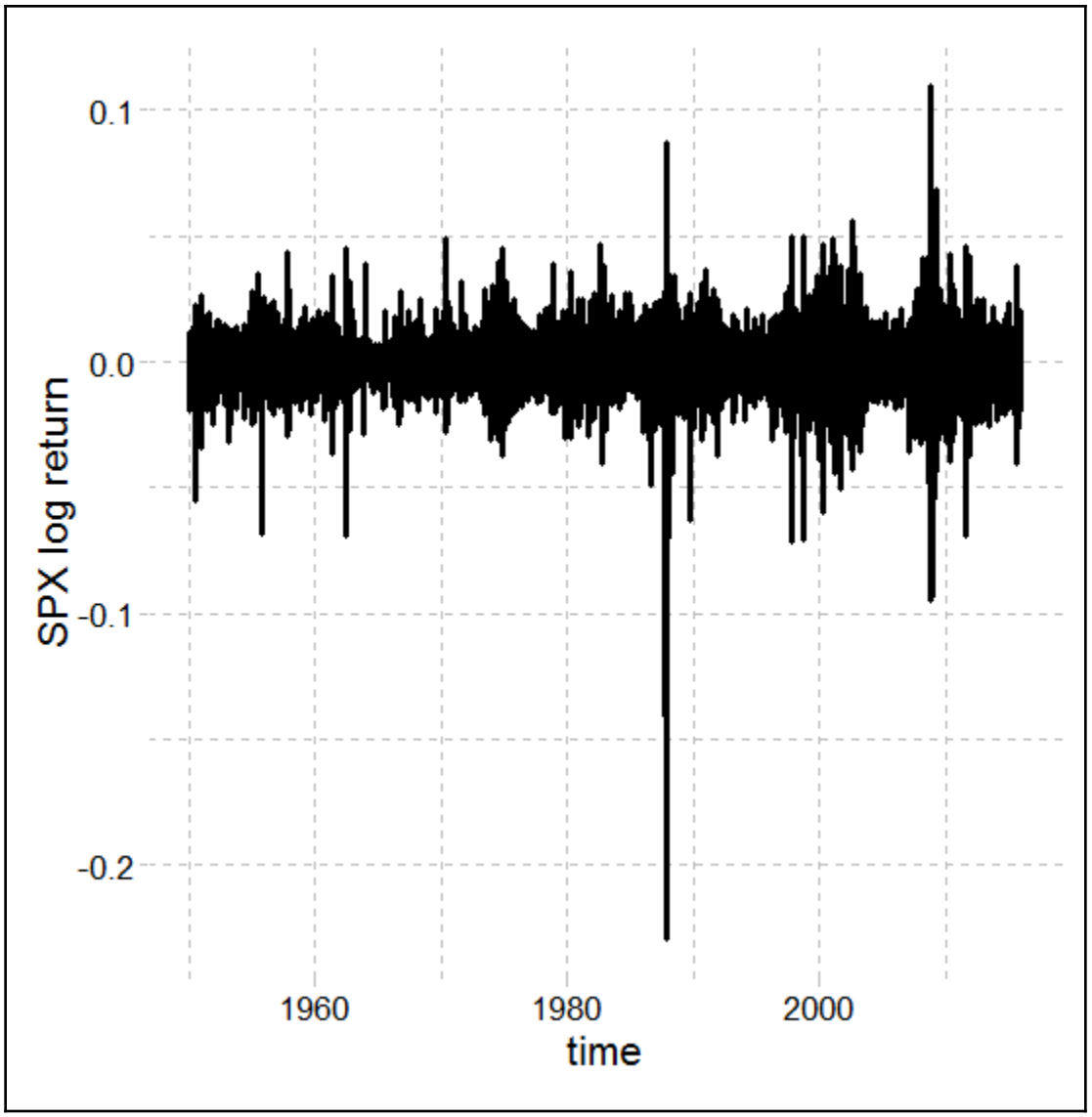
Total params: 1,350
Trainable params: 1,350
Non-trainable params: 0

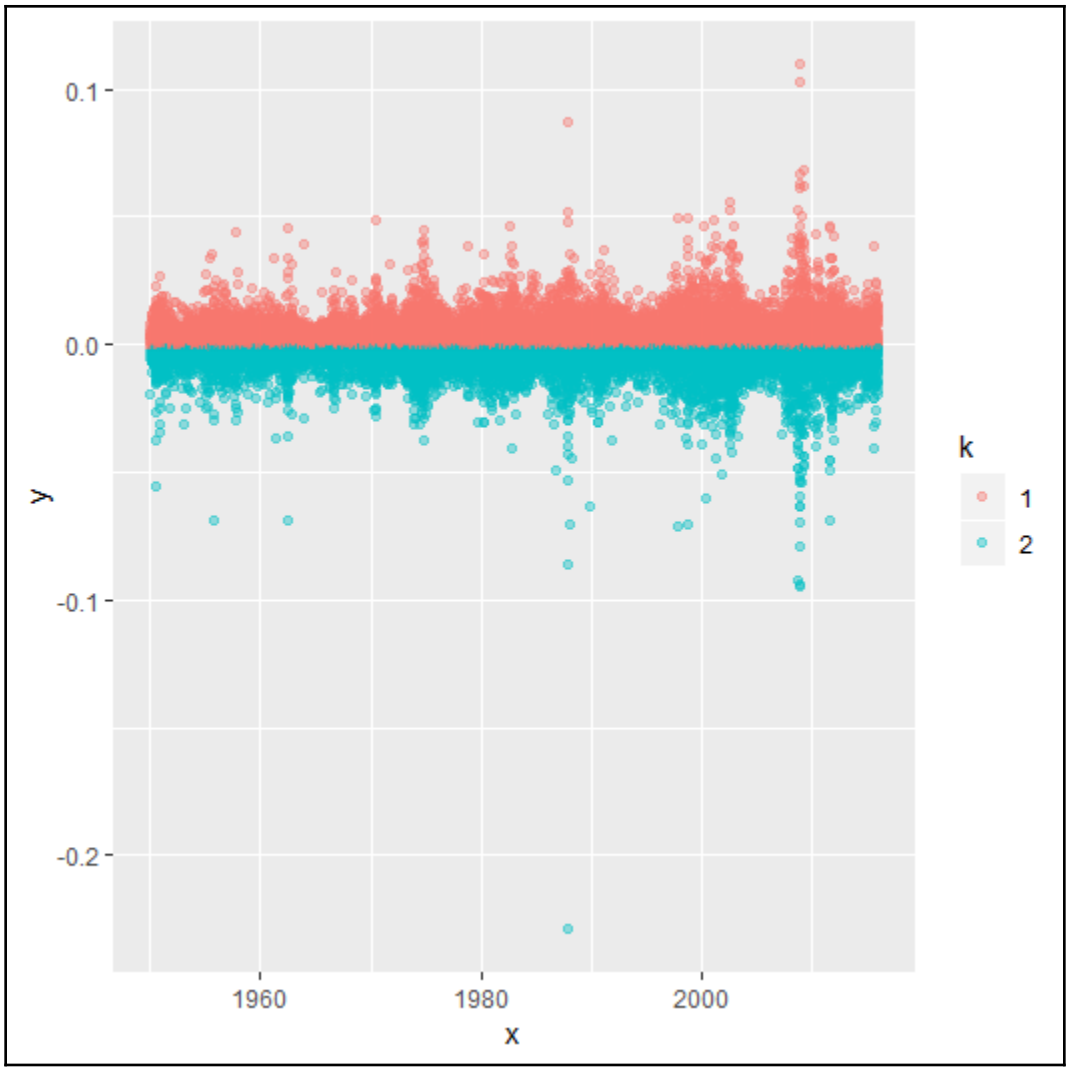

0.580
0.560
0.54
0.52
0.50
0.480
0.460
0.440
1
2
3
4
5
6
7
8
9
10
loss
val_loss
0.980
0.960
0.940
0.920
0.90
0.880
0.860
0.840
0.820
0.80
0.780
0.760
0.740
1
2
3
4
5
6
7
8
9
10
acc
val_acc

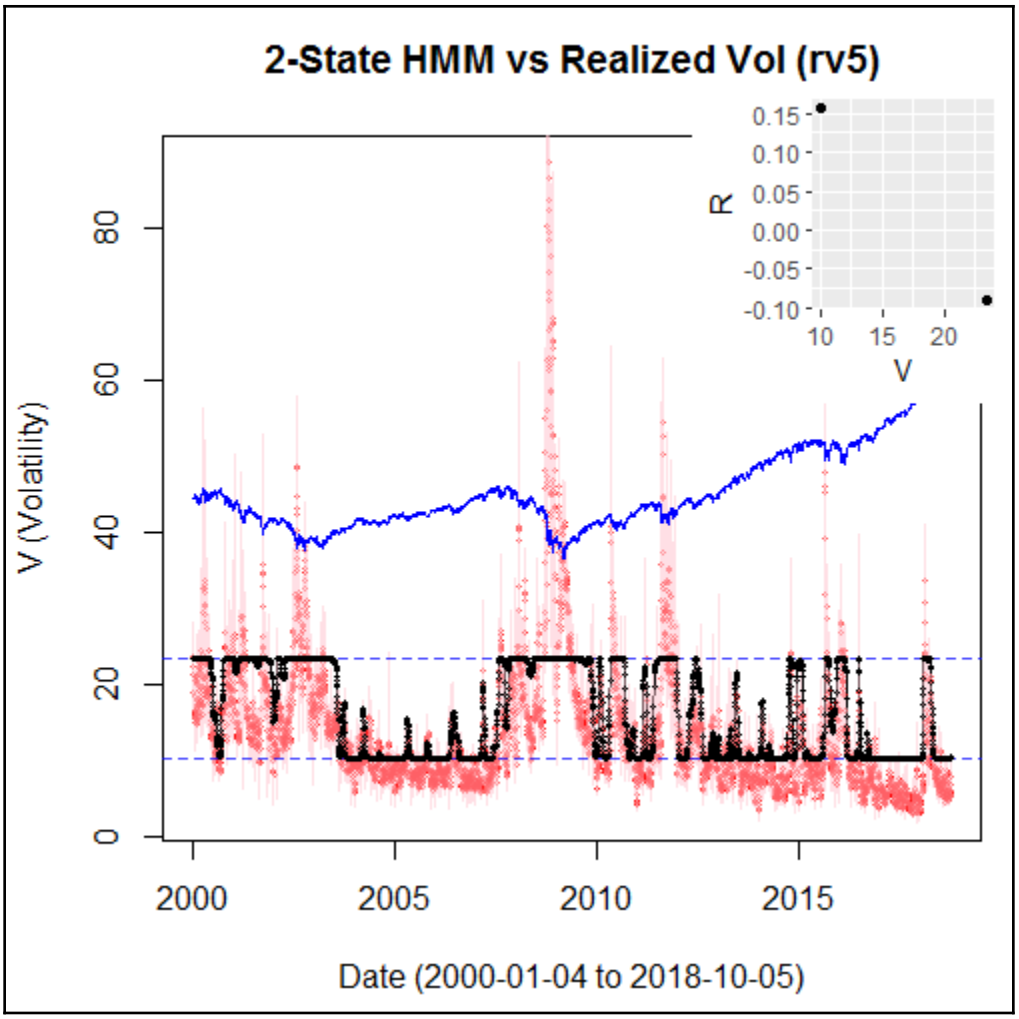
Chapter 9: Markovian in R



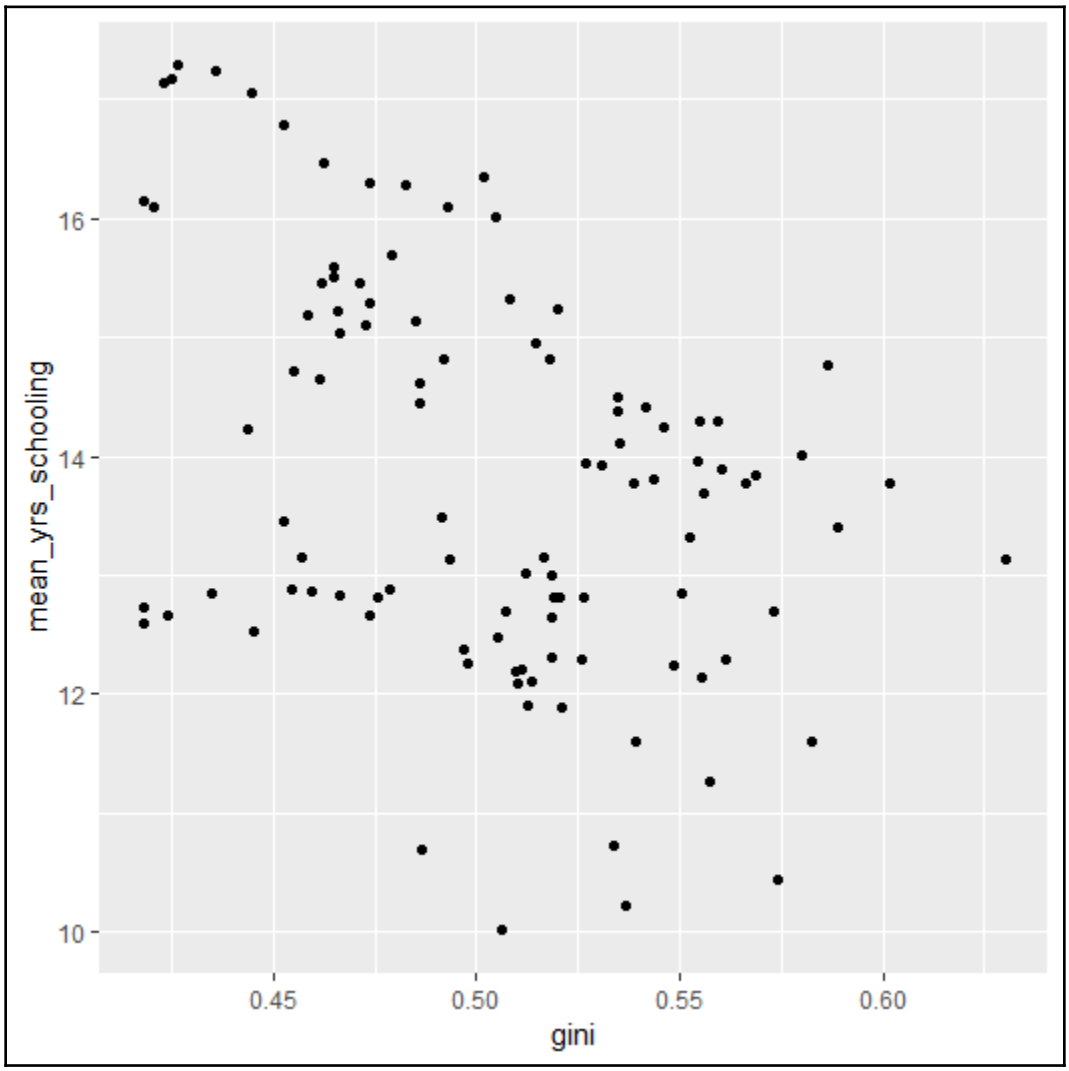

2000
1500
1000
500
0
SPX closing prices
1960
1980
2000
time

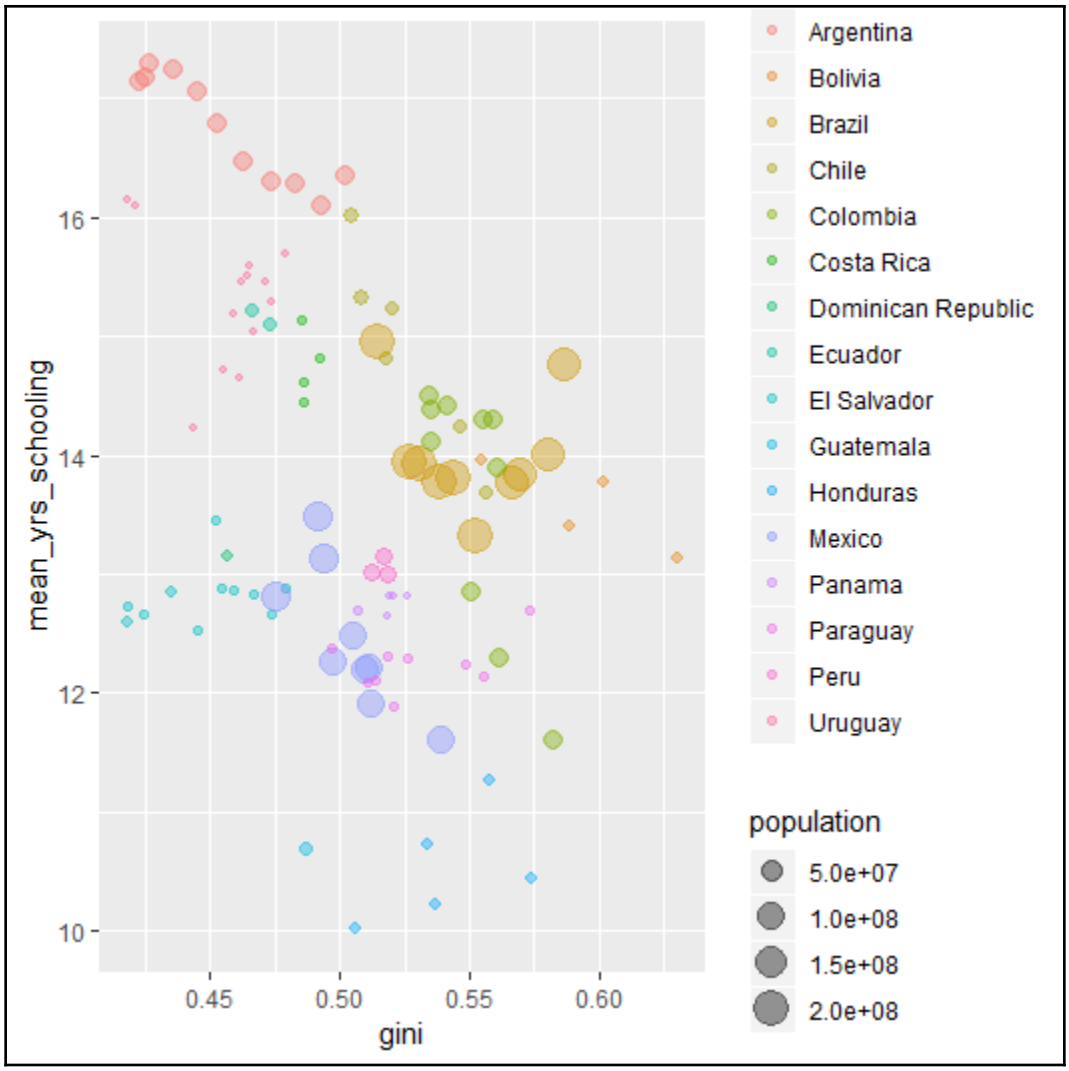

0.1
0.0
-0.1
-0.2
SPX log return
1960
1980
2000
time

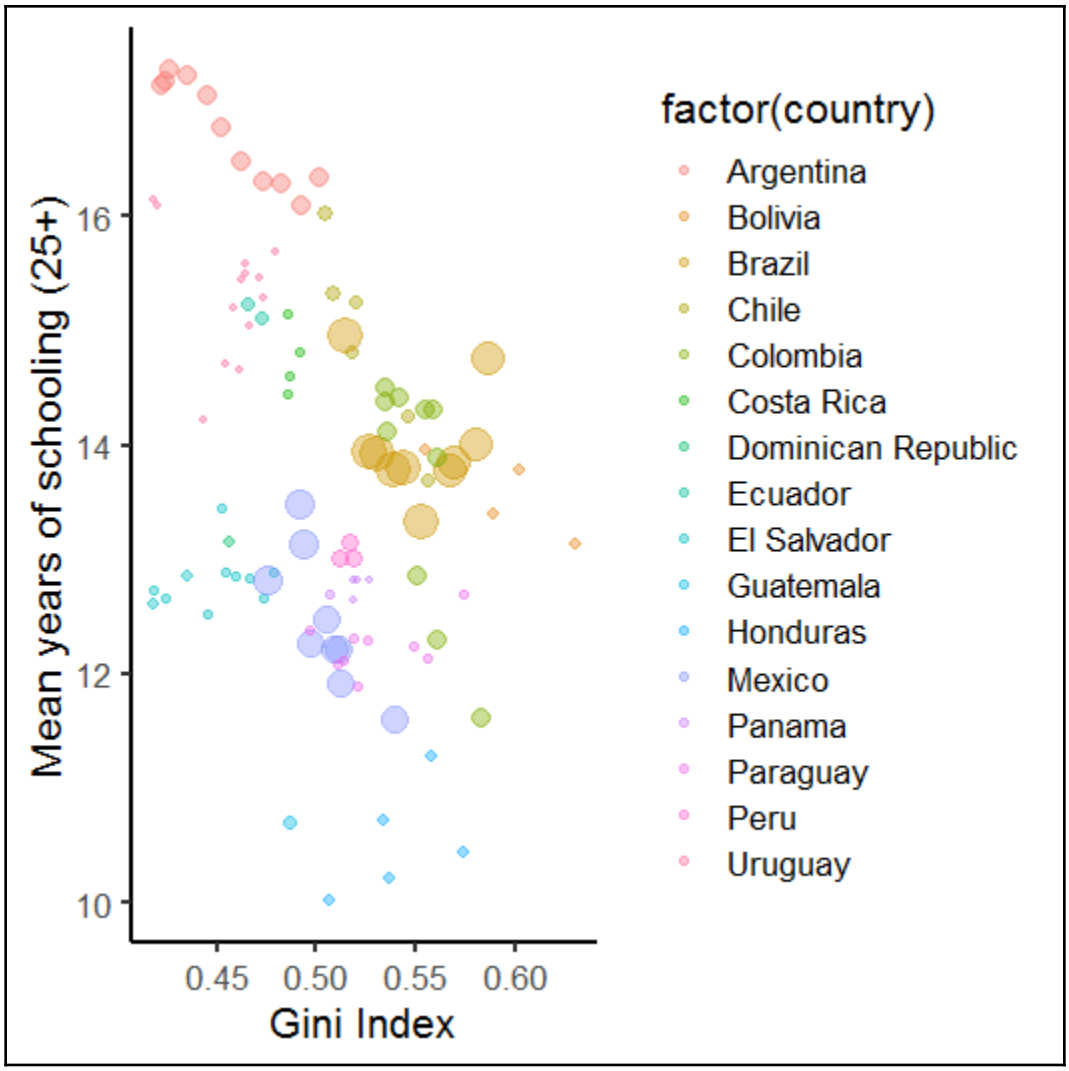

0.1
0.0
-0.1
-0.2
y
1960
1980
2000
x
k
1
2

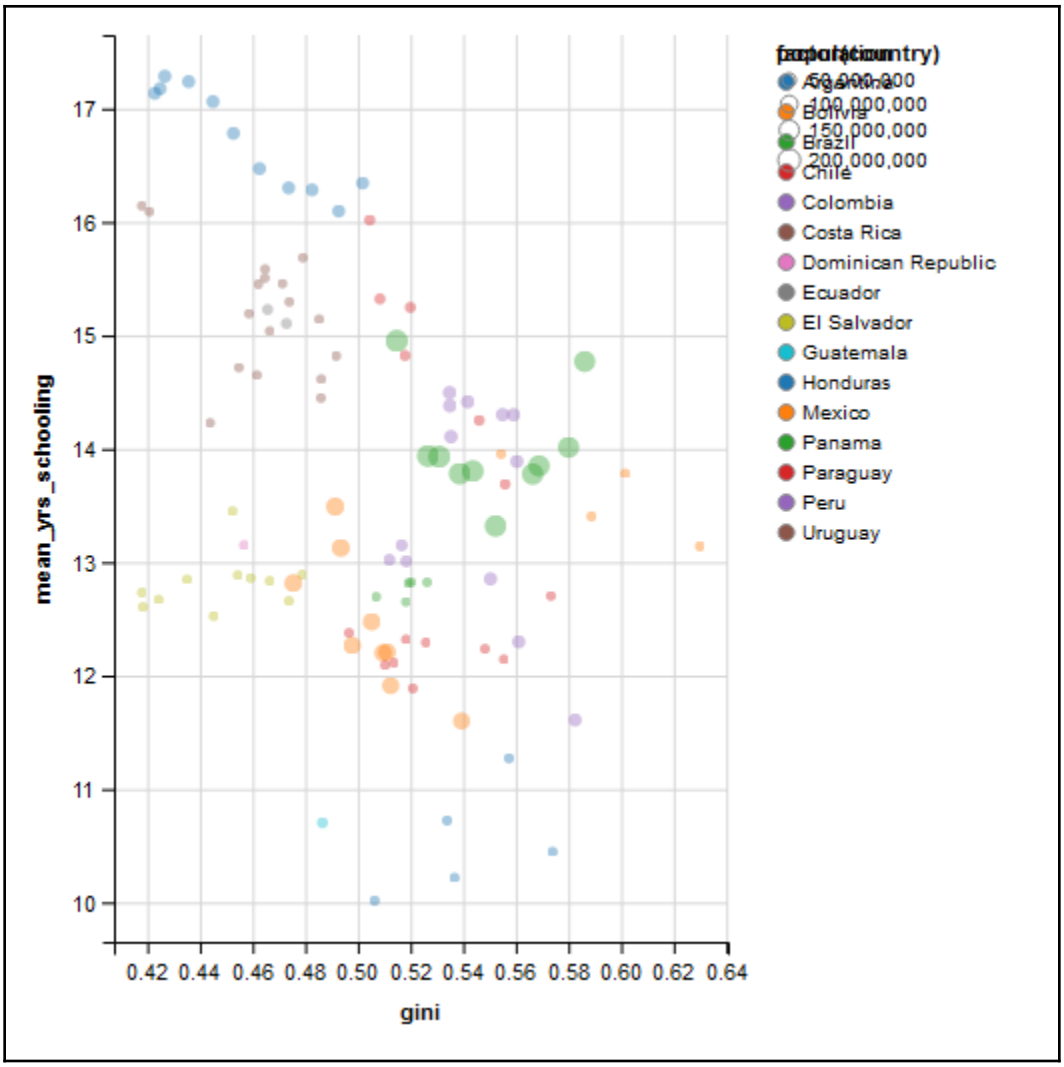

2-State HMM vs Realized Vol (rv5)
V (Volatility)
0
20
40
60
80
2000
2005
2010
2015
Date (2000-01-04 to 2018-10-05)
R
0.15
0.10
0.05
0.00
-0.05
-0.10
10
15
20
V

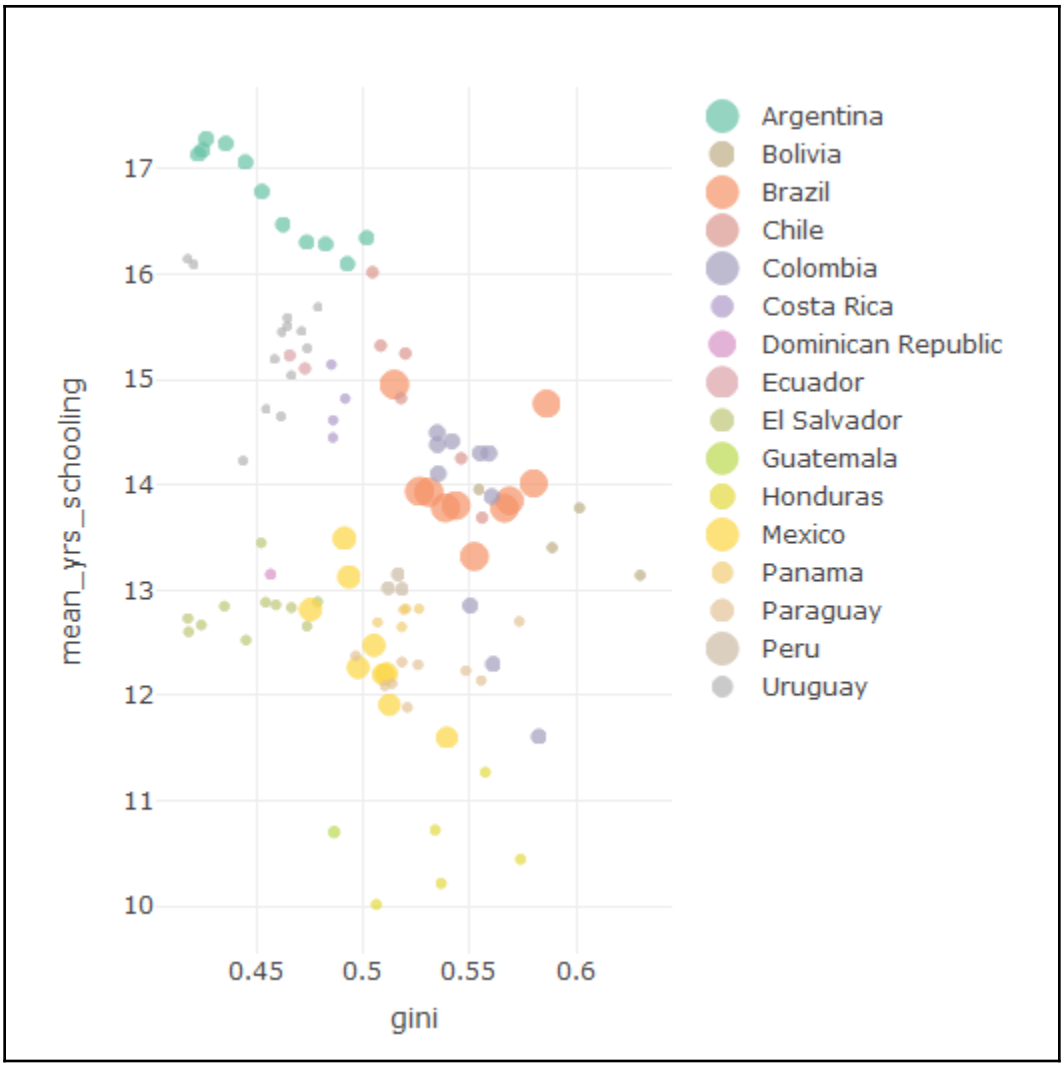
Chapter 10: Visualizing Data

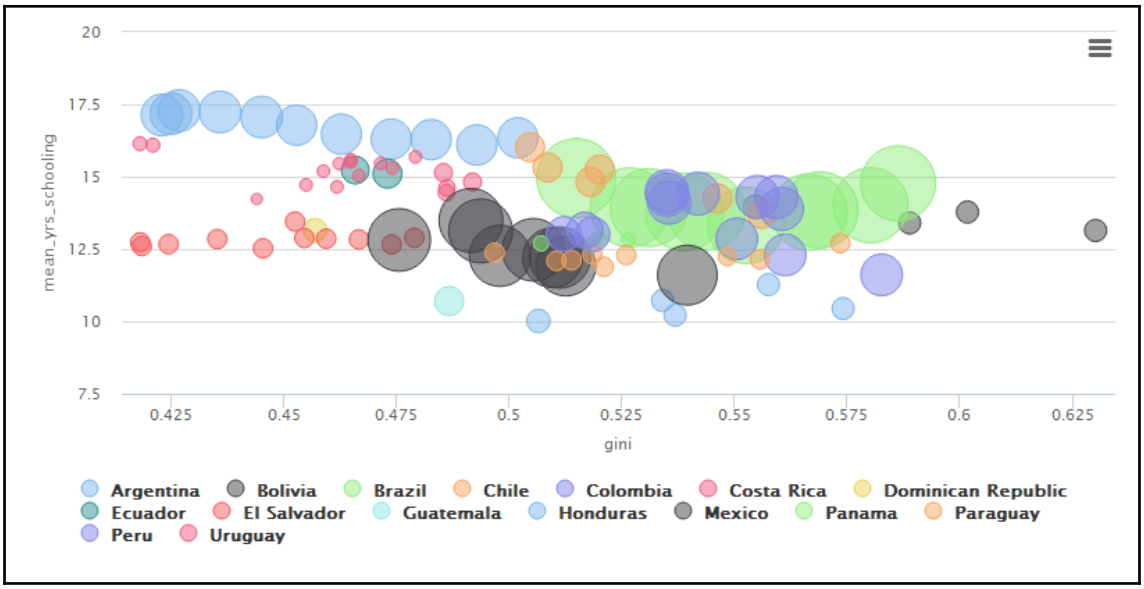


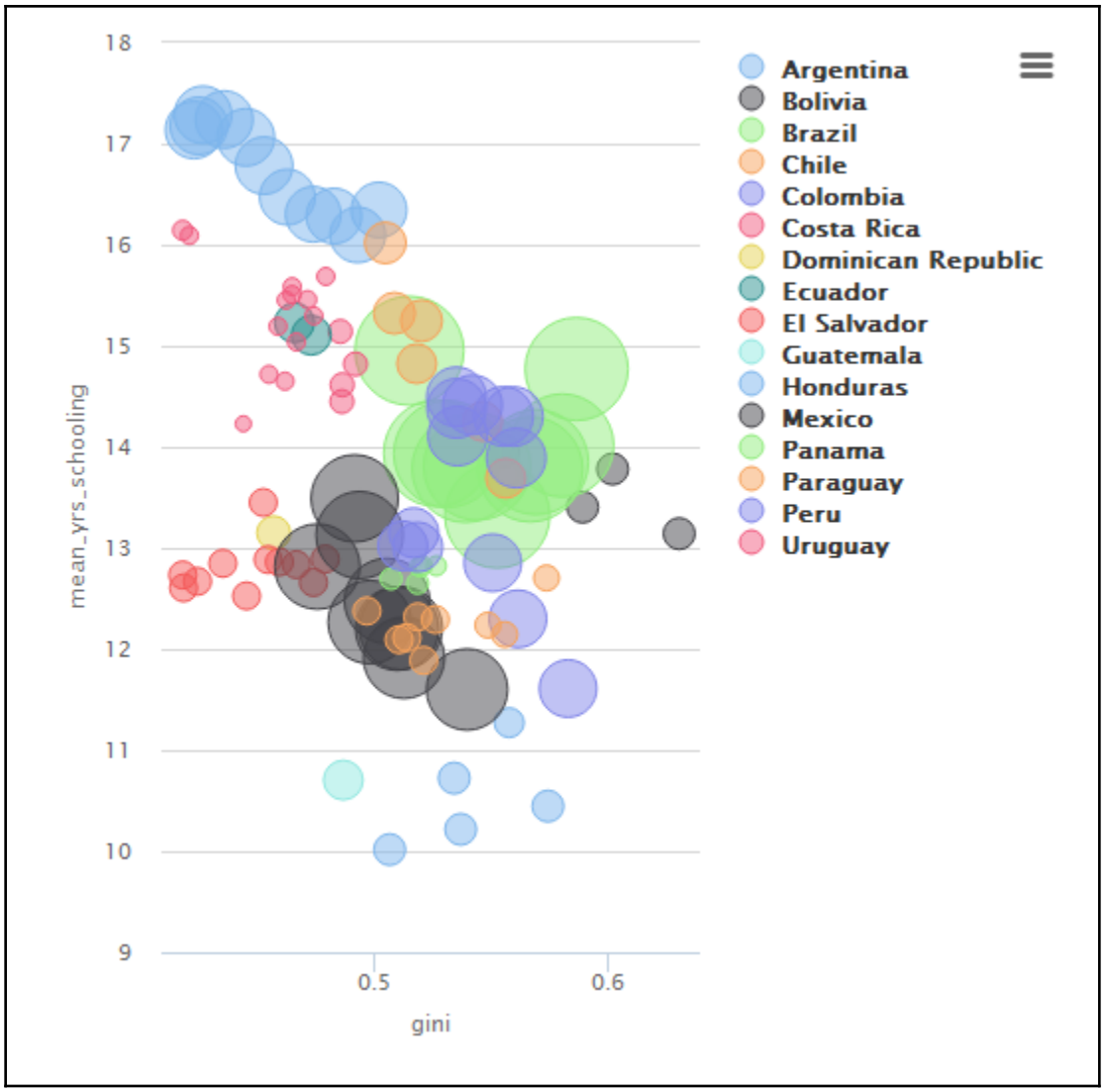

mean_yrs_schooling
16
14
12
10
0.45
0.50
0.55
0.60
gini
Argentina
Bolivia
Brazil
Chile
Colombia
Costa Rica
Dominican Republic
Ecuador
El Salvador
Guatemala
Honduras
Mexico
Panama
Paraguay
Peru
Uruguay
population
5.0e+07
1.0e+08
1.5e+08
2.0e+08

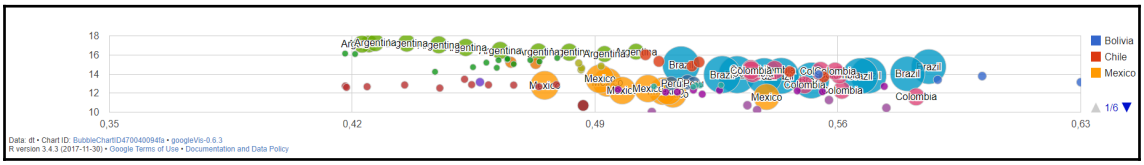

Mean years of schooling (25+)
16
14
12
10
0.45
0.50
0.55
0.60
Gini Index
factor(country)
Argentina
Bolivia
Brazil
Chile
Colombia
Costa Rica
Dominican Republic
Ecuador
El Salvador
Guatemala
Honduras
Mexico
Panama
Paraguay
Peru
Uruguay

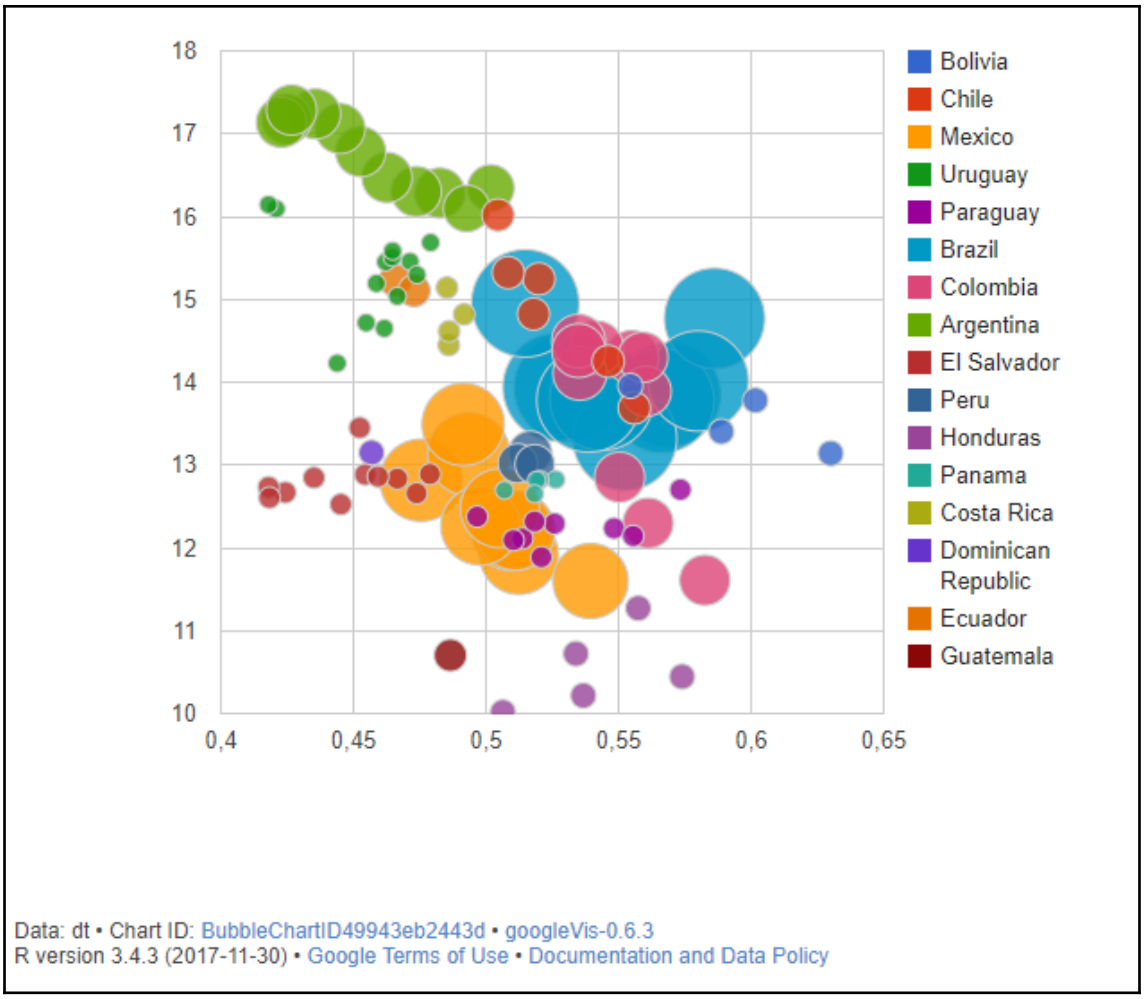

mean_yrs_schooling
17
16
15
14
13
12
11
10
0.42 0.44 0.46 0.48 0.50 0.52 0.54 0.56 0.58 0.60 0.62 0.64
gini
Argentina
Bolivia
Brazil
Chile
Colombia
Costa Rica
Dominican Republic
Ecuador
El Salvador
Guatemala
Honduras
Mexico
Panama
Paraguay
Peru
Uruguay

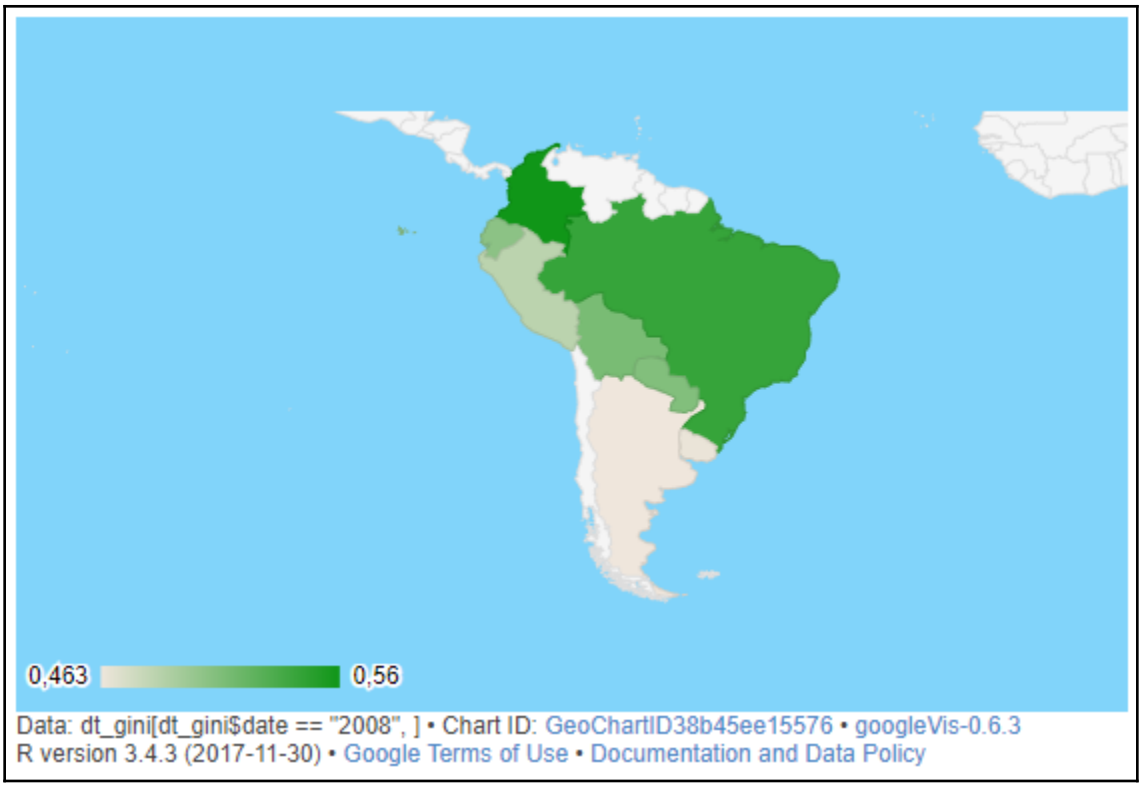

mean_yrs_schooling
17
16
15
14
13
12
11
10
0.45
0.5
0.55
0.6
gini
Argentina
Bolivia
Brazil
Chile
Colombia
Costa Rica
Dominican Republic
Ecuador
El Salvador
Guatemala
Honduras
Mexico
Panama
Paraguay
Peru
Uruguay


20
17.5
15
12.5
10
7.5
mean_yrs_schooling
0.425
0.45
0.475
0.5
0.525
0.55
0.575
0.6
0.625
gini
Argentina
Bolivia
Brazil
Chile
Colombia
Costa Rica
Dominican Republic
Ecuador
El Salvador
Guatemala
Honduras
Mexico
Panama
Paraguay
Peru
Uruguay

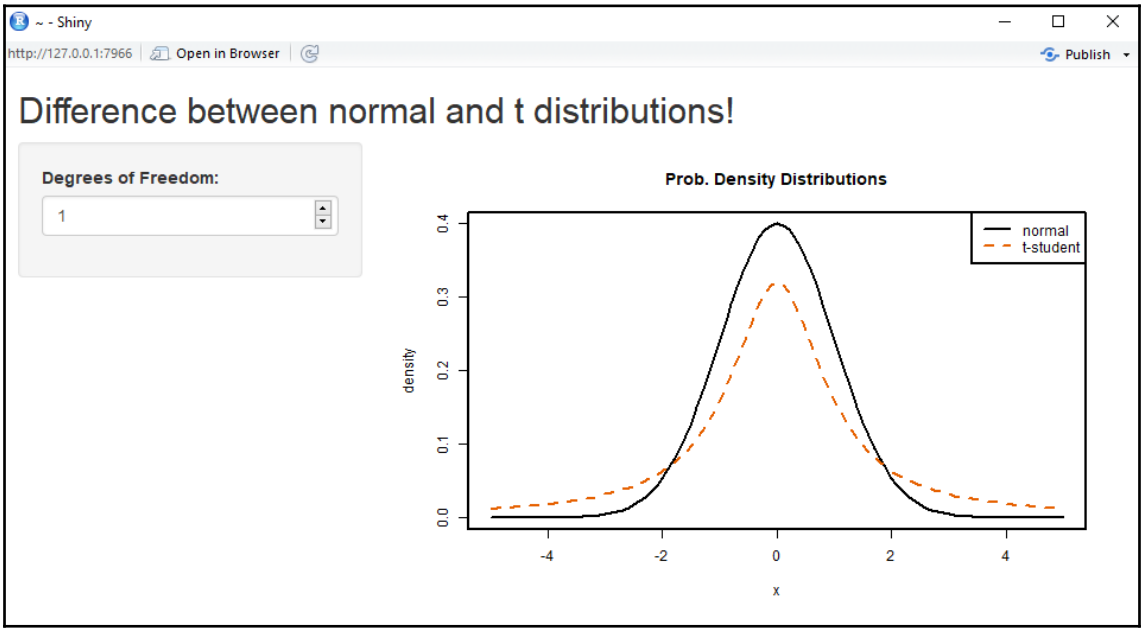

mean_yrs_schooling
gini
18
17
16
15
14
13
12
11
10
9
0.5
0.6
Argentina
Bolivia
Brazil
Chile
Colombia
Costa Rica
Dominican Republic
Ecuador
El Salvador
Guatemala
Honduras
Mexico
Panama
Paraguay
Peru
Uruguay

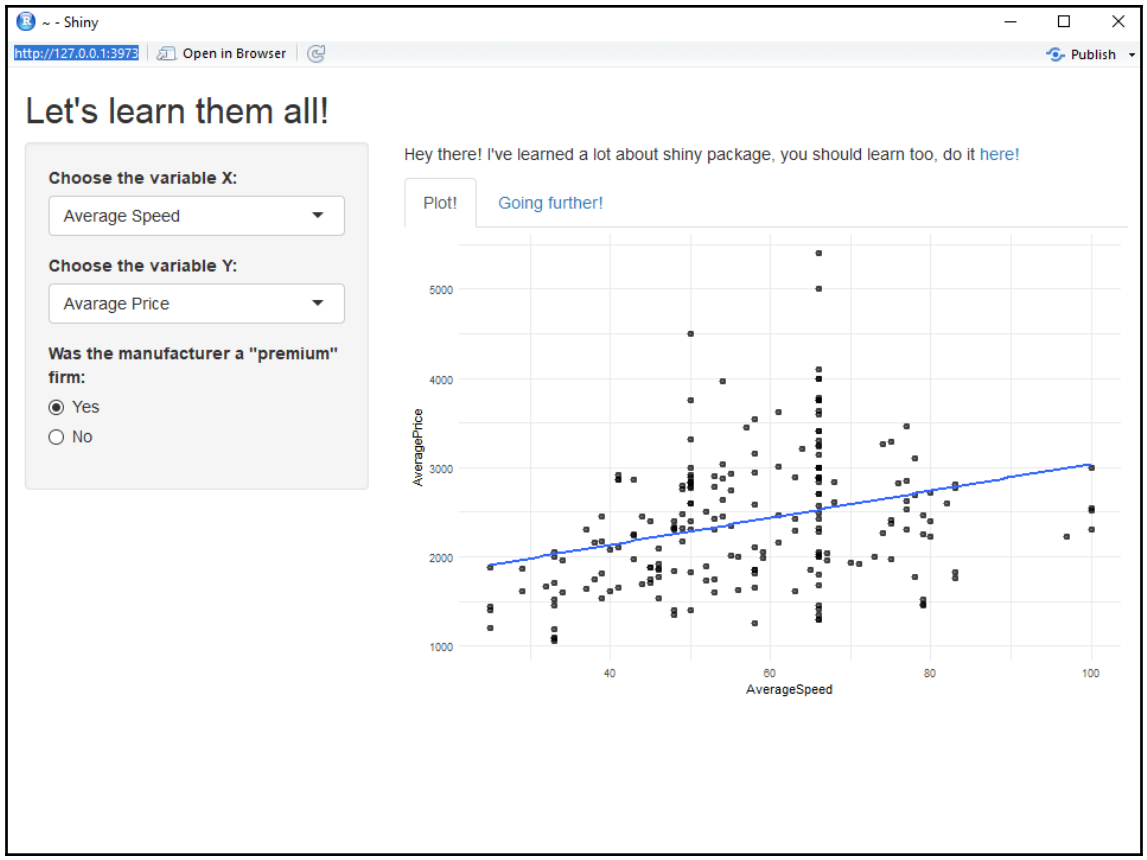

18
16
14
12
10
0,35
0,42
0,49
0,56
0,63
Bolivia
Chile
Mexico
1/6
Data: dt • Chart ID: BubbleChartID470400941a • googleVis-0.6.3
R version 3.4.3 (2017-11-30) • Google Terms of Use • Documentation and Data Policy


18
17
16
15
14
13
12
11
10
0,4
0,45
0,5
0,55
0,6
0,65
Bolivia
Chile
Mexico
Uruguay
Paraguay
Brazil
Colombia
Argentina
El Salvador
Peru
Honduras
Panama
Costa Rica
Dominican Republic
Ecuador
Guatemala
Data: dt • Chart ID: BubbleChartID49943eb2443d • googleVis-0.6.3
R version 3.4.3 (2017-11-30) • Google Terms of Use • Documentation and Data Policy

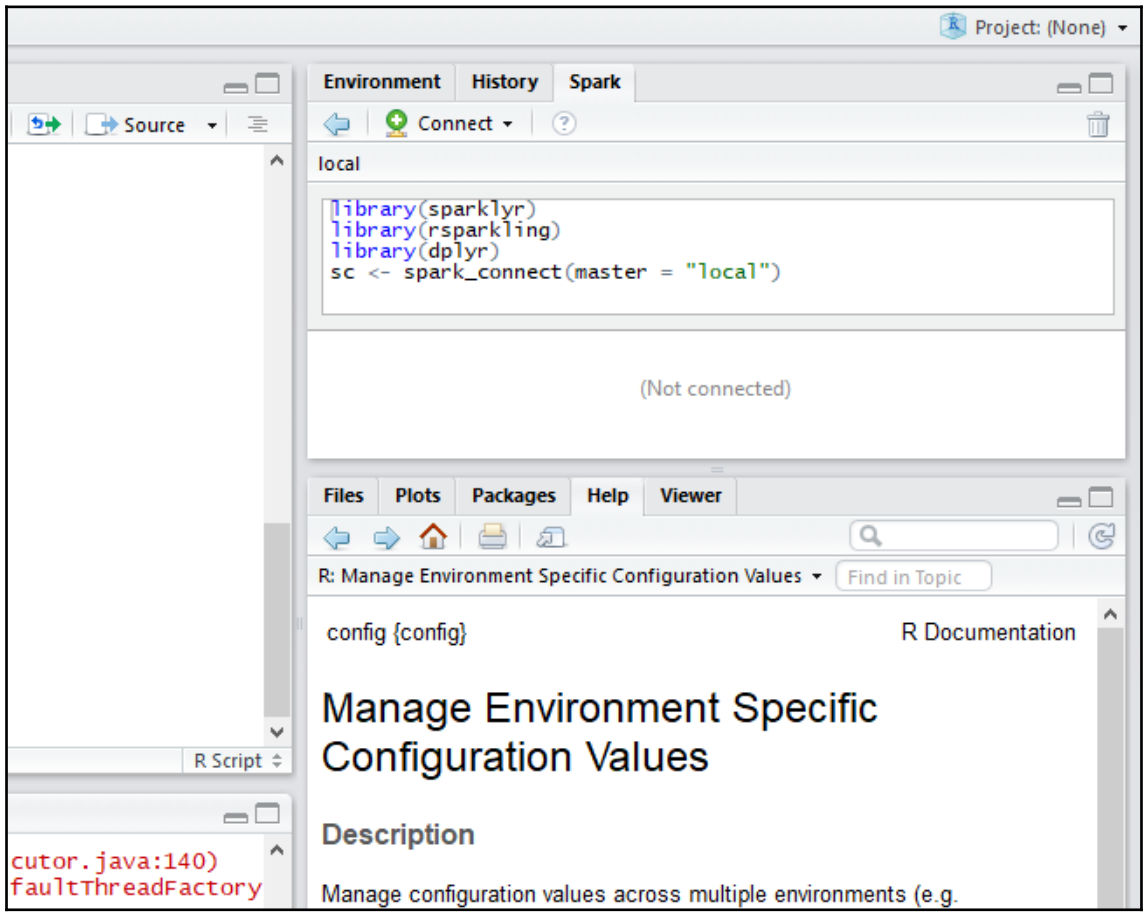

0,463
0,56
Data: dt_gini[dt_gini$date == "2008",] • Chart ID: GeoChartID38b45ee15576 • googleVis-0.6.3
R version 3.4.3 (2017-11-30) • Google Terms of Use • Documentation and Data Policy

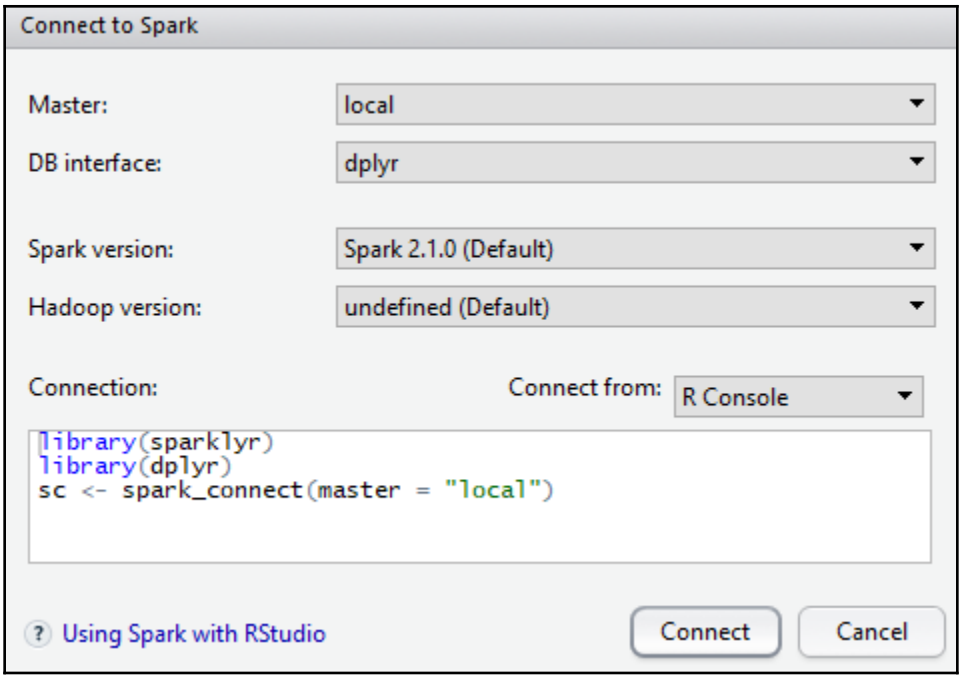
Chapter 11: Going to Production with R



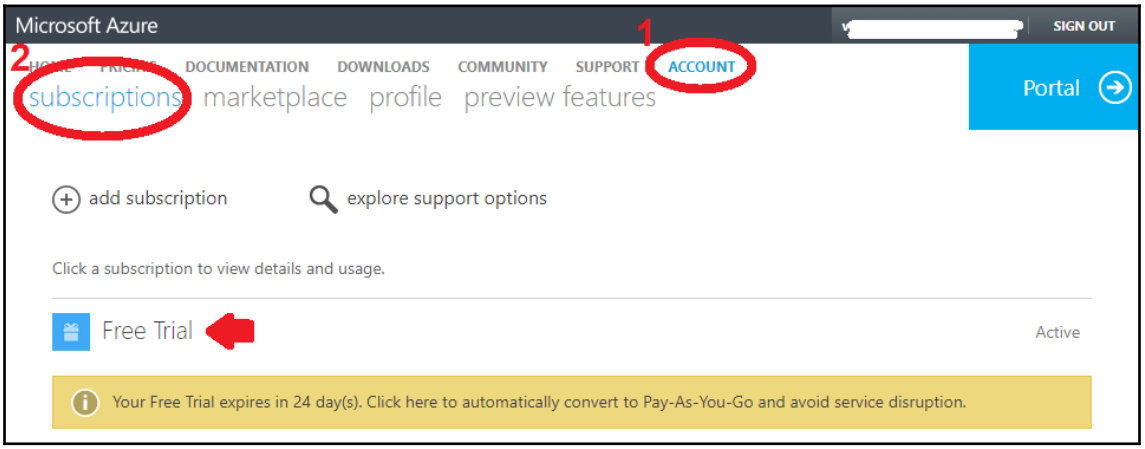

~ - Shiny
http://127.0.0.1:3973
Open in Browser
Publish
Let's learn them all!
Choose the variable X:
Average Speed
Choose the variable Y:
Avarage Price
Was the manufacturer a "premium" firm:
Yes
No
Hey there! I've learned a lot about shiny package, you should learn too, do it here!
Plot!
Going further!
AverageSpeed
AverageTrice
5000
4000
3000
2000
1000
40
60
80
100

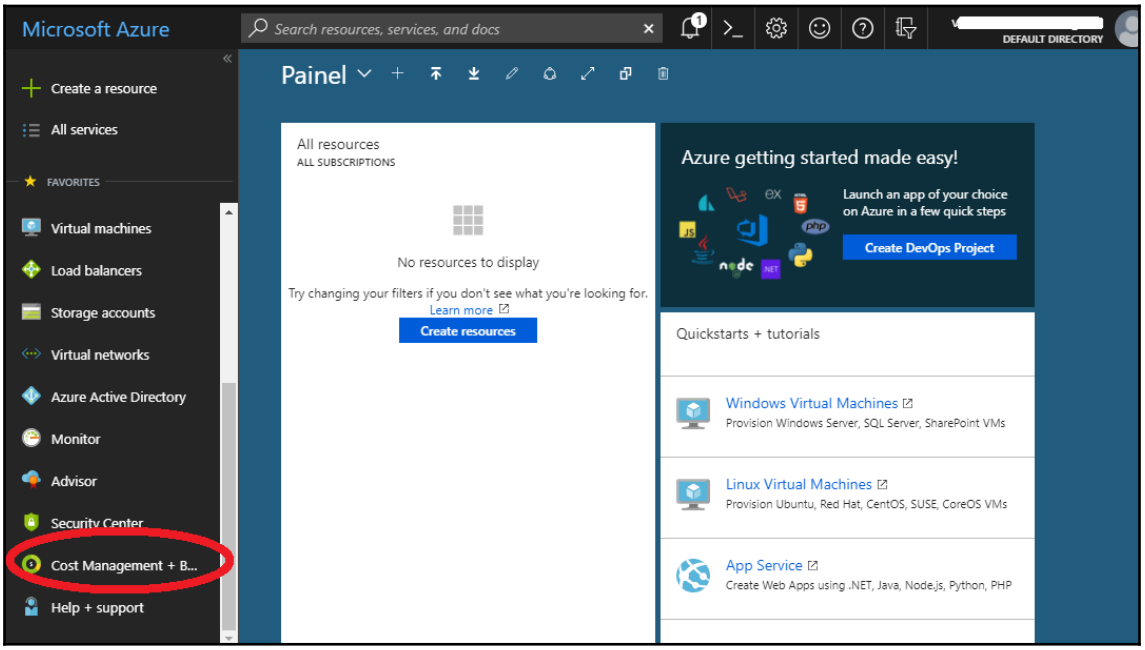
Chapter 12: Large Scale Data Analytics with Hadoop

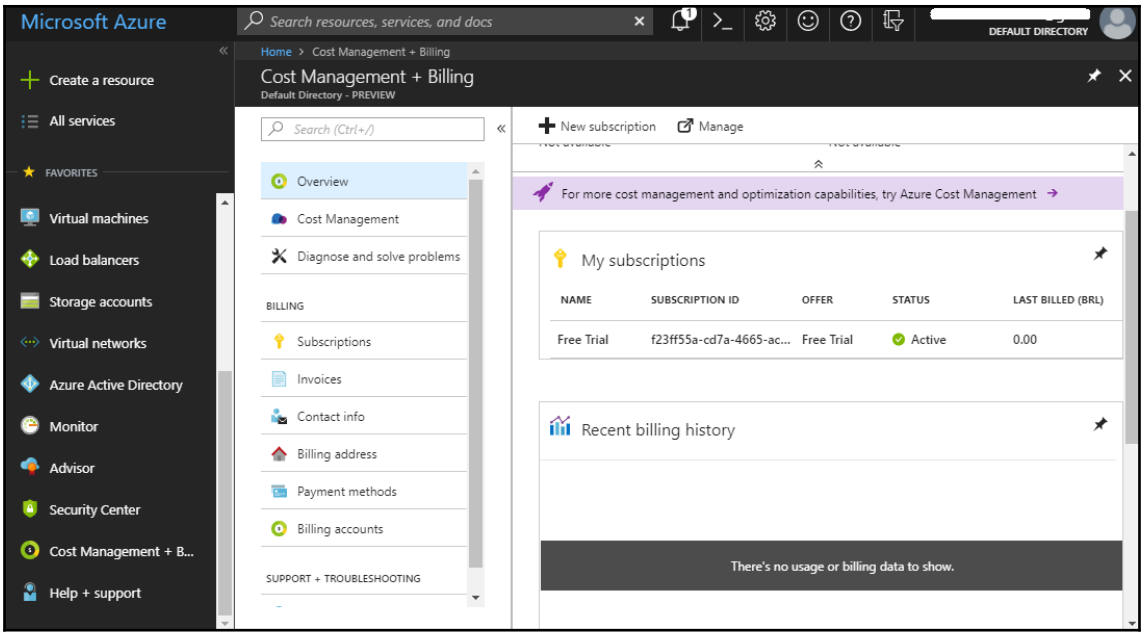


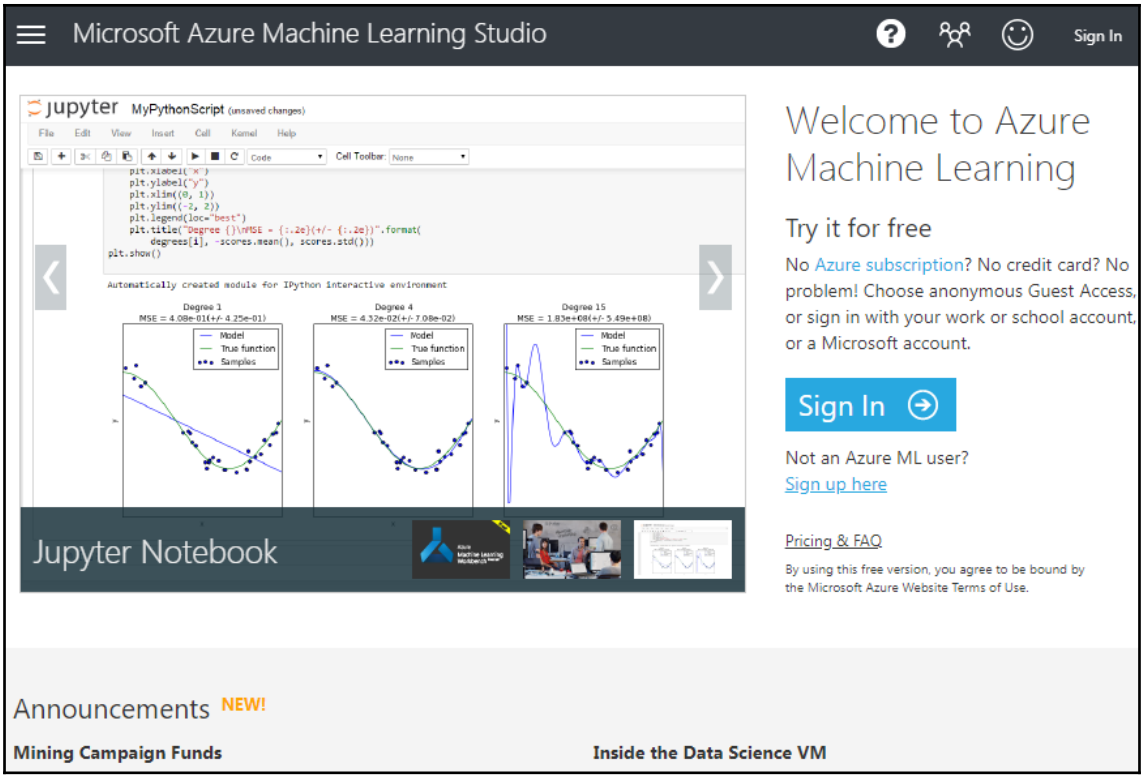

Connect to Spark
Master:
local
DB interface:
dplyr
Spark version:
Spark 2.1.0 (Default)
Hadoop version:
undefined (Default)
Connection:
Connect from:
R Console
library(sparklyr)
library(dplyr)
sc <- spark_connect(master = "local")
Using Spark with RStudio
Connect
Cancel

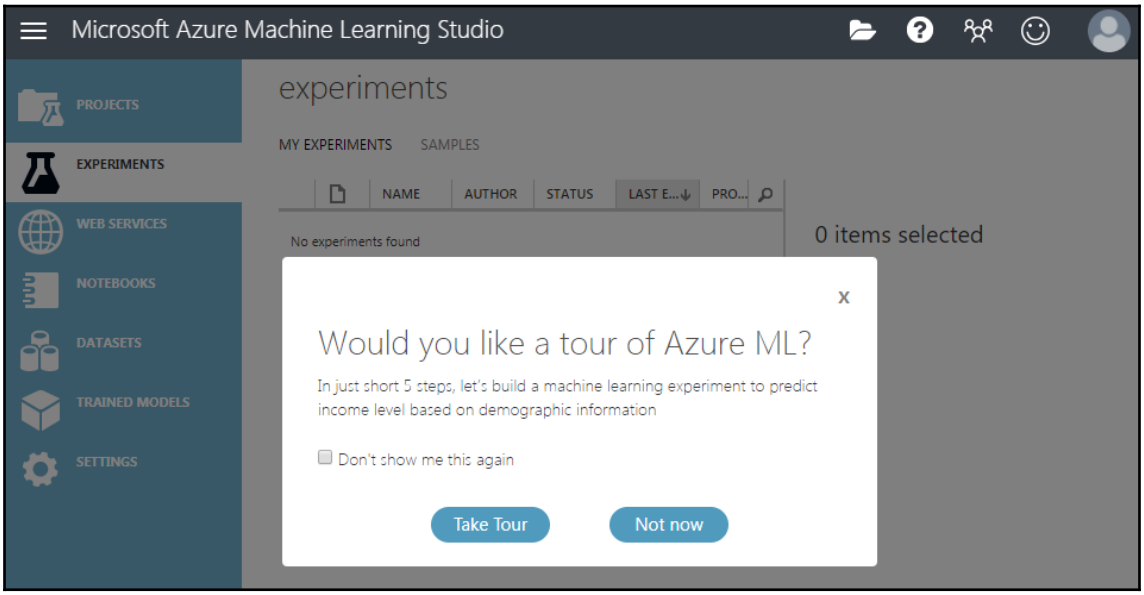
Chapter 13: R on Cloud

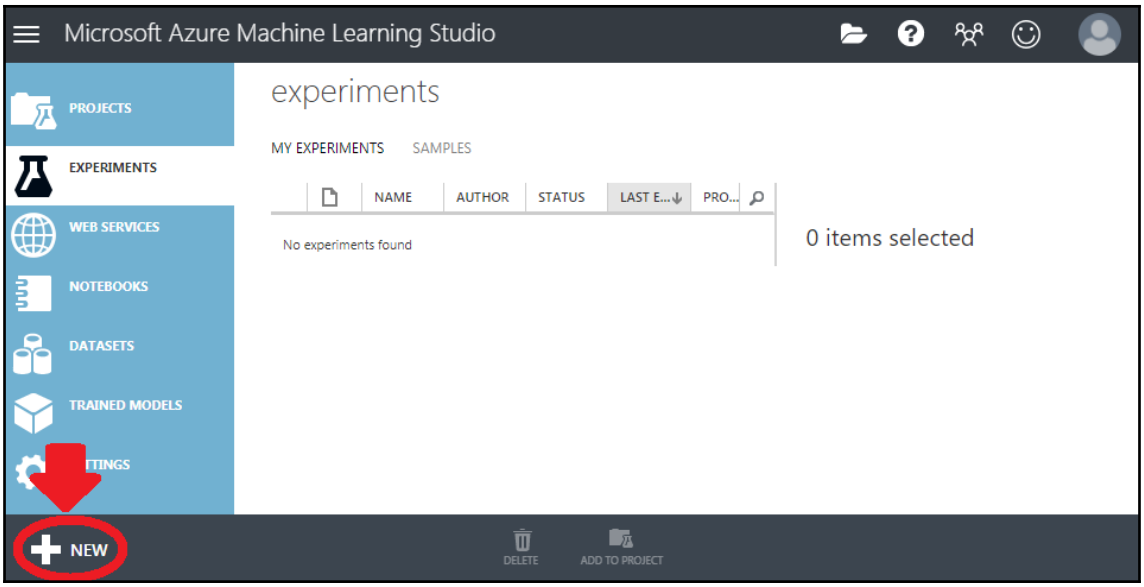


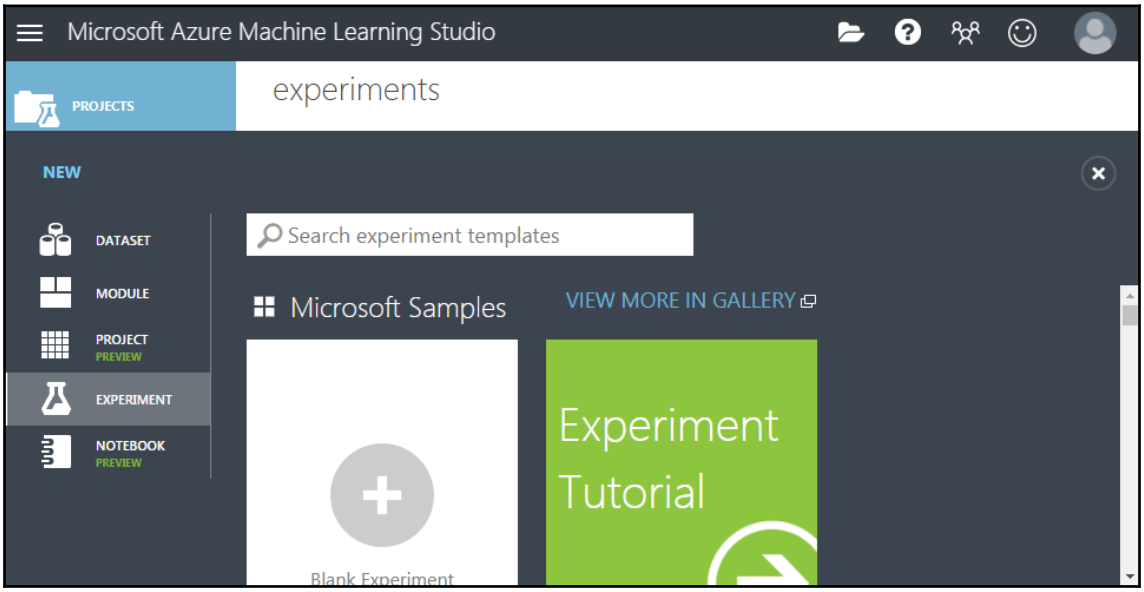


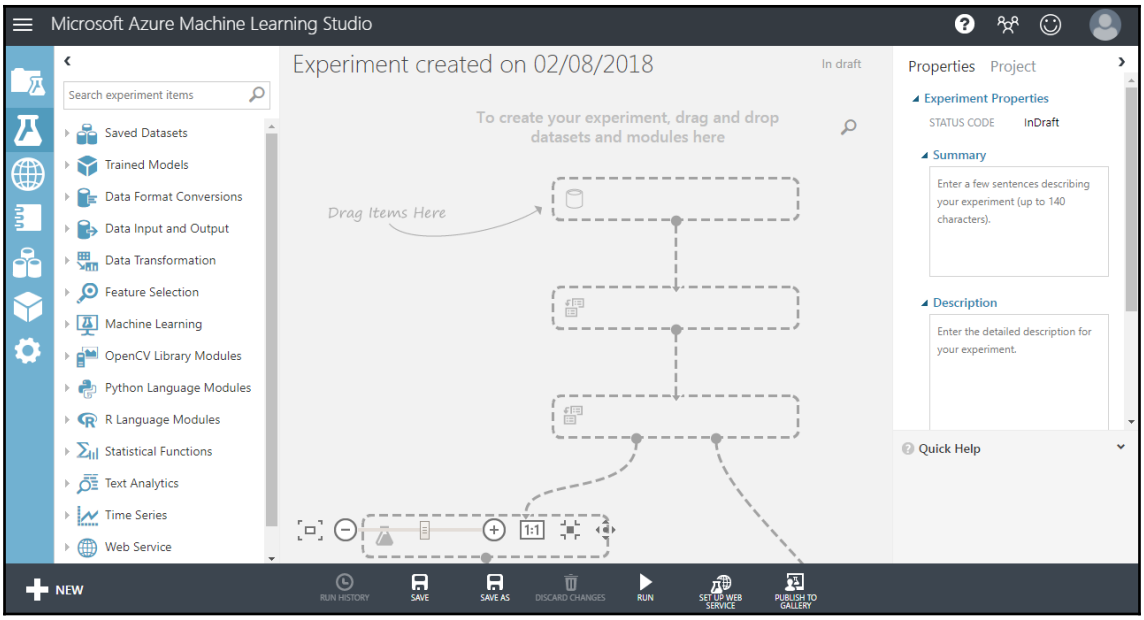

Microsoft Azure
Search resources, services, and docs
DEFAULT DIRECTORY
Create a resource
All services
FAVORITES
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Security Center
Cost Management + B...
Help + support
Home > Cost Management + Billing
Cost Management + Billing
Default Directory - PREVIEW
Search (Ctrl+/)
Overview
Cost Management
Diagnose and solve problems
BILLING
Subscriptions
Invoices
Contact info
Billing address
Payment methods
Billing accounts
SUPPORT + TROUBLESHOOTING
New subscription
Manage
For more cost management and optimization capabilities, try Azure Cost Management
My subscriptions
NAME
SUBSCRIPTION ID
OFFER
STATUS
LAST BILLED (BRL)
Free Trial
f23ff55a-cd7a-4665-ac...
Free Trial
Active
0.00
Recent billing history
There's no usage or billing data to show.

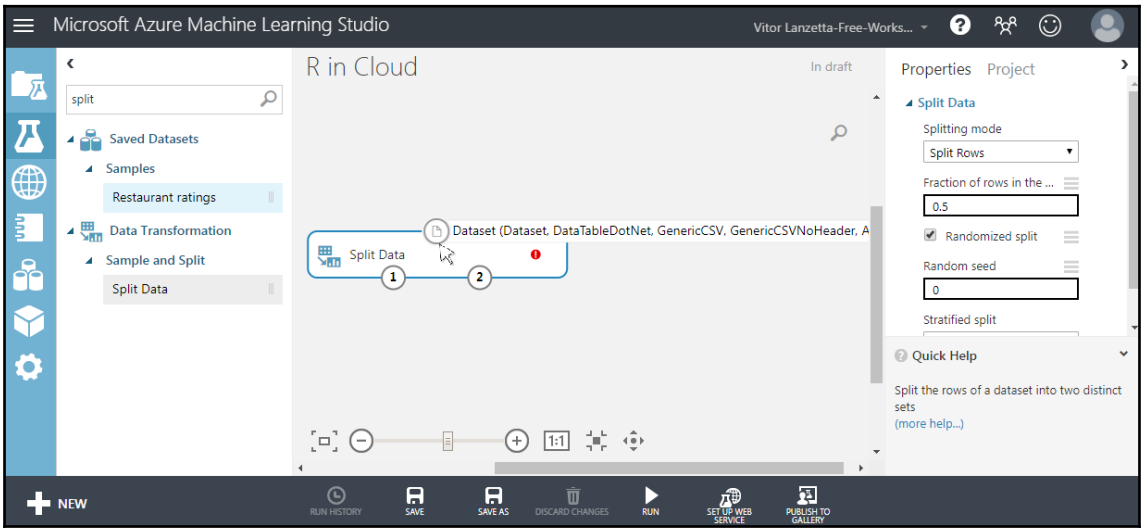

Microsoft Azure Machine Learning Studio
Sign In
jupyter MyPythonScript (unsaved changes)
Jupyter Notebook
Welcome to Azure Machine Learning
Try it for free
No Azure subscription? No credit card? No problem! Choose anonymous Guest Access, or sign in with your work or school account, or a Microsoft account.
Sign In
Not an Azure ML user?
Sign up here
Pricing & FAQ
By using this free version, you agree to be bound by the Microsoft Azure Website Terms of Use.
Announcements NEW!
Mining Campaign Funds
Inside the Data Science VM

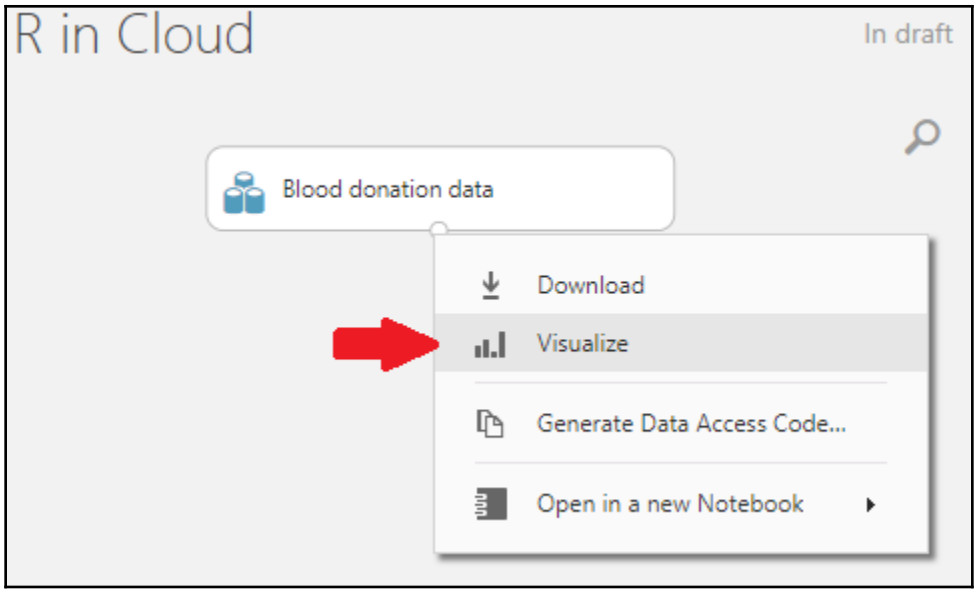

Microsoft Azure Machine Learning Studio
PROJECTS
EXPERIMENTS
WEB SERVICES
NOTEBOOKS
DATASETS
TRAINED MODELS
SETTINGS
experiments
MY EXPERIMENTS
SAMPLES
NAME
AUTHOR
STATUS
LAST E...
PRO...
No experiments found
0 items selected
Would you like a tour of Azure ML?
In just short 5 steps, let's build a machine learning experiment to predict income level based on demographic information
Don't show me this again
Take Tour
Not now

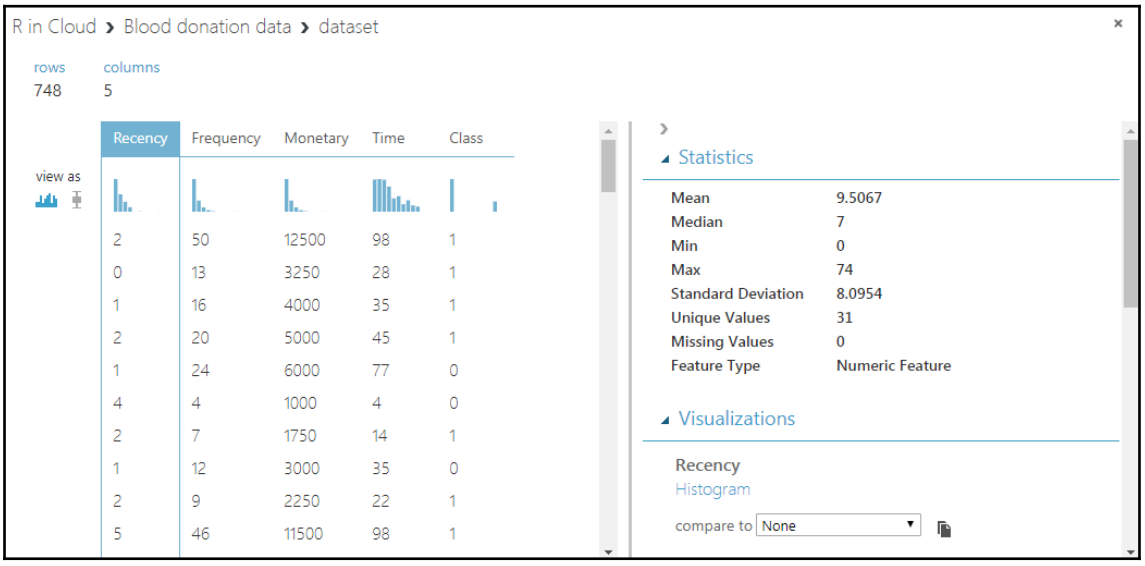

Microsoft Azure Machine Learning Studio
PROJECTS
EXPERIMENTS
WEB SERVICES
NOTEBOOKS
DATASETS
TRAINED MODELS
experiments
MY EXPERIMENTS
SAMPLES
NAME
AUTHOR
STATUS
LAST E...
PRO...
No experiments found
0 items selected
NEW
DELETE
ADD TO PROJECT

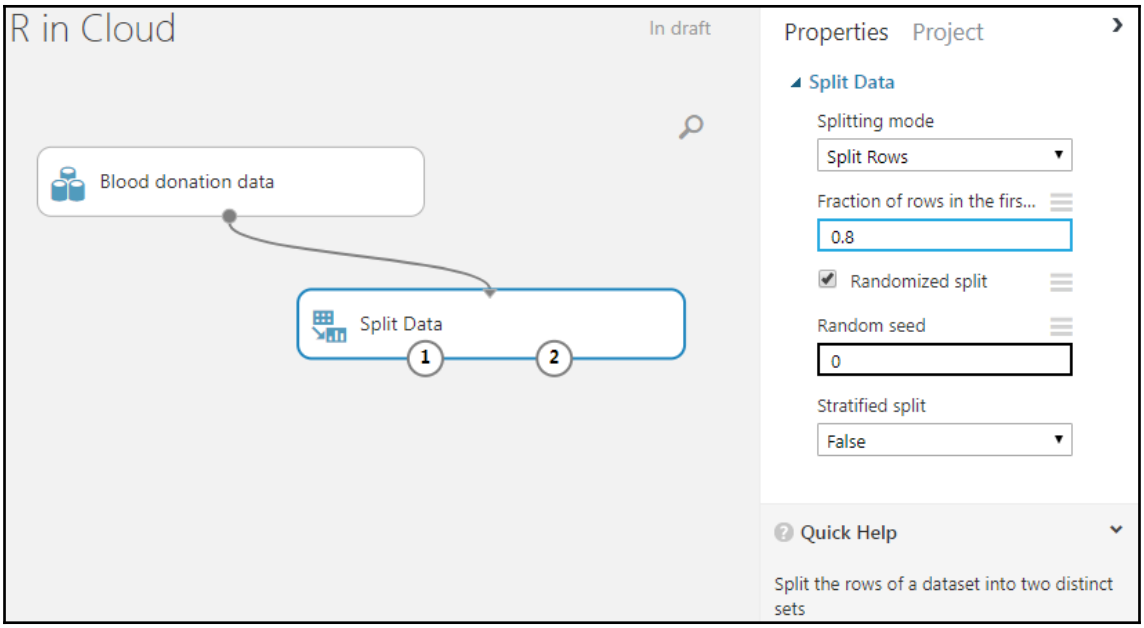

Microsoft Azure Machine Learning Studio
PROJECTS
experiments
NEW
DATASET
MODULE
PROJECT
PREVIEW
EXPERIMENT
NOTEBOOK
PREVIEW
Search experiment templates
Microsoft Samples
VIEW MORE IN GALLERY
Blank Experiment
Experiment
Tutorial

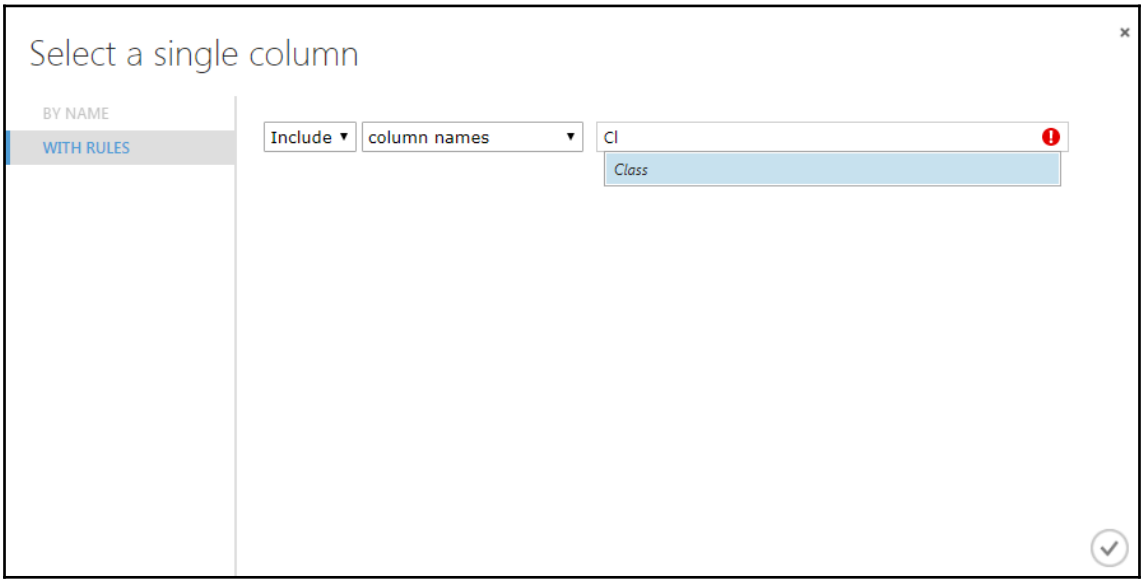

Microsoft Azure Machine Learning Studio
Experiment created on 02/08/2018
In draft
Search experiment items
Saved Datasets
Trained Models
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Time Series
Web Service
To create your experiment, drag and drop datasets and modules here
Drag Items Here
Properties
Project
Experiment Properties
STATUS CODE
InDraft
Summary
Enter a few sentences describing your experiment (up to 140 characters).
Description
Enter the detailed description for your experiment.
Quick Help
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

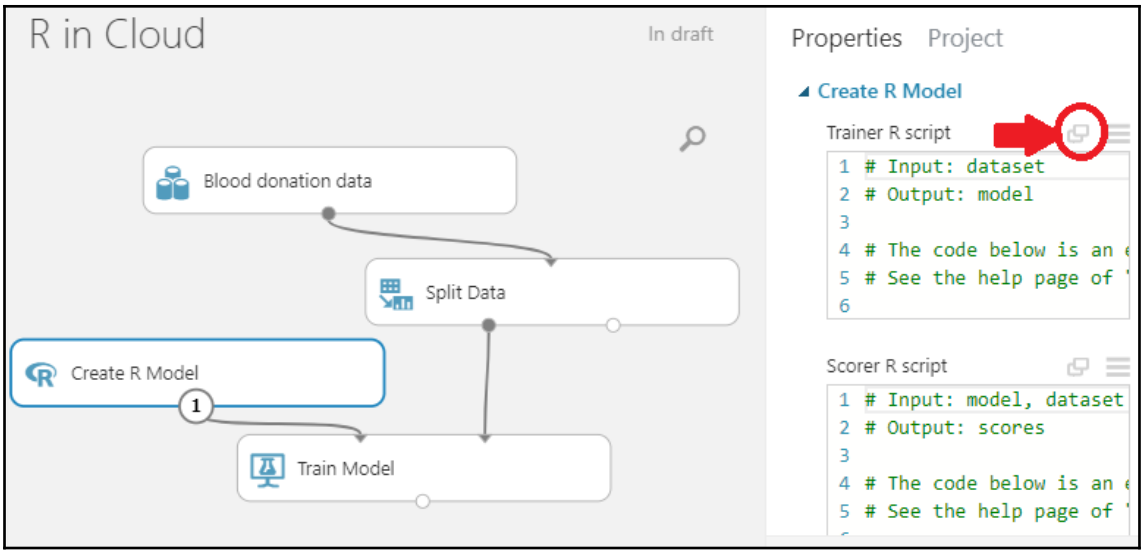

Microsoft Azure Machine Learning Studio
Vitor Lanzetta-Free-Works...
R in Cloud
In draft
split
Saved Datasets
Samples
Restaurant ratings
Data Transformation
Sample and Split
Split Data
Dataset (Dataset, DataTableDotNet, GenericCSV, GenericCSVNoHeader, A
Split Data
Properties
Project
Split Data
Splitting mode
Split Rows
Fraction of rows in the ...
0.5
Randomized split
Random seed
0
Stratified split
Quick Help
Split the rows of a dataset into two distinct sets
(more help...)
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

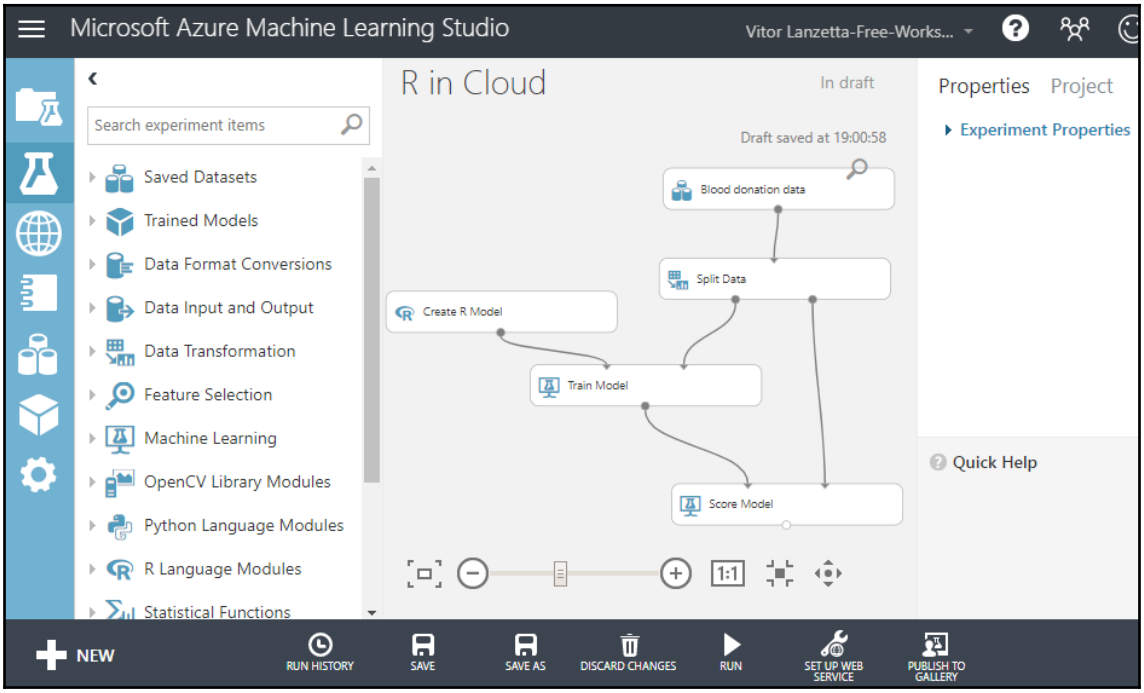

R in Cloud
In draft
Blood donation data
Download
Visualize
Generate Data Access Code...
Open in a new Notebook

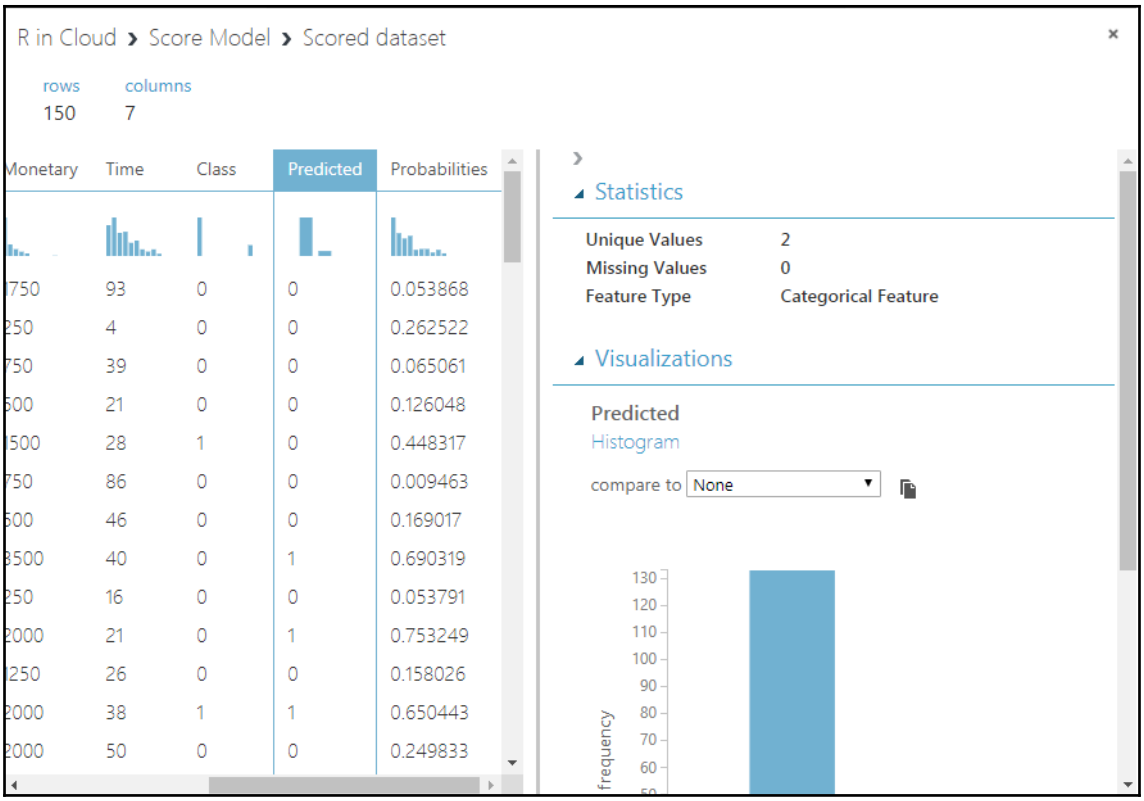

R in Cloud › Blood donation data › dataset
rows
748
columns
5
view as
Recency Frequency Monetary Time Class
2 50 12500 98 1
0 13 3250 28 1
1 16 4000 35 1
2 20 5000 45 1
1 24 6000 77 0
4 4 1000 4 0
2 7 1750 14 1
1 12 3000 35 0
2 9 2250 22 1
5 46 11500 98 1
Statistics
Mean 9.5067
Median 7
Min 0
Max 74
Standard Deviation 8.0954
Unique Values 31
Missing Values 0
Feature Type Numeric Feature
Visualizations
Recency
Histogram
compare to None

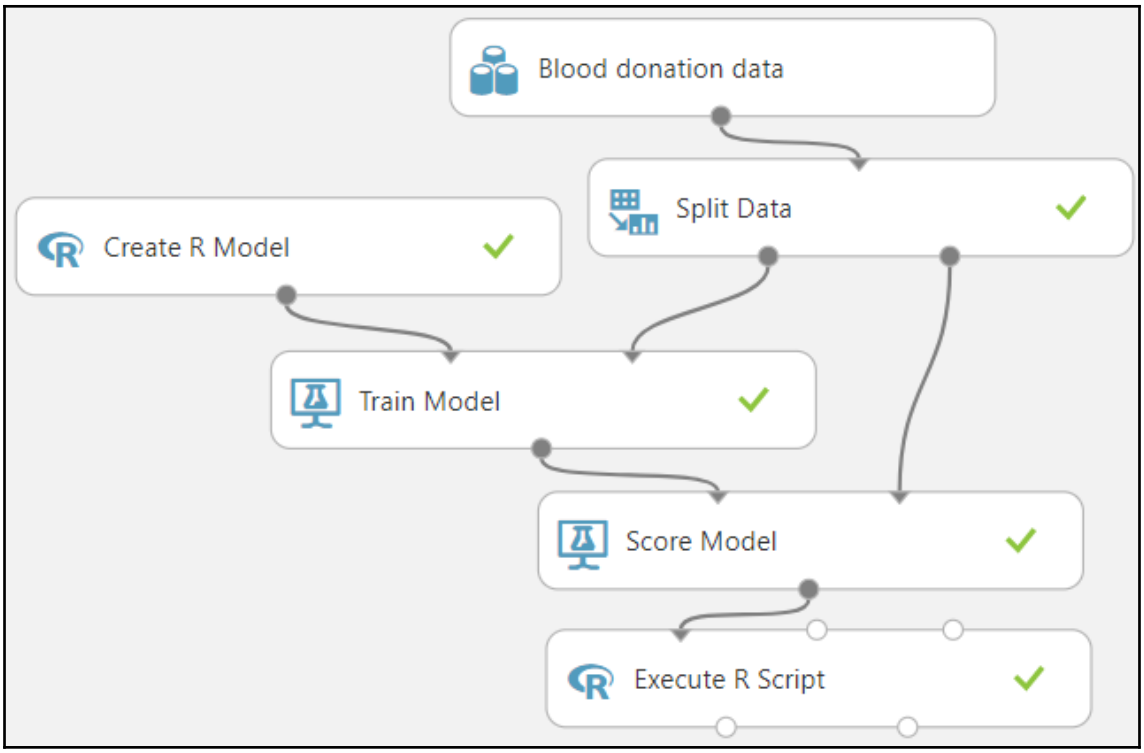

R in Cloud
In draft
Blood donation data
Split Data
1
2
Properties
Project
Split Data
Splitting mode
Split Rows
Fraction of rows in the firs...
0.8
Randomized split
Random seed
0
Stratified split
False
Quick Help
Split the rows of a dataset into two distinct sets


Select a single column
BY NAME
WITH RULES
Include
column names
Cl
Class


R in Cloud
In draft
Properties
Project
Create R Model
Trainer R script
1 # Input: dataset
2 # Output: model
3
4 # The code below is an
5 # See the help page of
6
Scorer R script
1 # Input: model, dataset
2 # Output: scores
3
4 # The code below is an
5 # See the help page of
Blood donation data
Split Data
Create R Model
1
Train Model


Microsoft Azure Machine Learning Studio
Vitor Lanzetta-Free-Works...
Search experiment items
Saved Datasets
Trained Models
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
R in Cloud
In draft
Draft saved at 19:00:58
Blood donation data
Split Data
Create R Model
Train Model
Score Model
1:1
Properties
Project
Experiment Properties
Quick Help
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY


R in Cloud › Score Model › Scored dataset
rows
150
columns
7
Monetary
Time
Class
Predicted
Probabilities
1750 93 0 0 0.053868
250 4 0 0 0.262522
750 39 0 0 0.065061
500 21 0 0 0.126048
1500 28 1 0 0.448317
750 86 0 0 0.009463
500 46 0 0 0.169017
3500 40 0 1 0.690319
250 16 0 0 0.053791
2000 21 0 1 0.753249
1250 26 0 0 0.158026
2000 38 1 1 0.650443
2000 50 0 0 0.249833
Statistics
Unique Values 2
Missing Values 0
Feature Type Categorical Feature
Visualizations
Predicted
Histogram
compare to None
130
120
110
100
90
80
70
60
frequency


Blood donation data
Split Data
Create R Model
Train Model
Score Model
Execute R Script