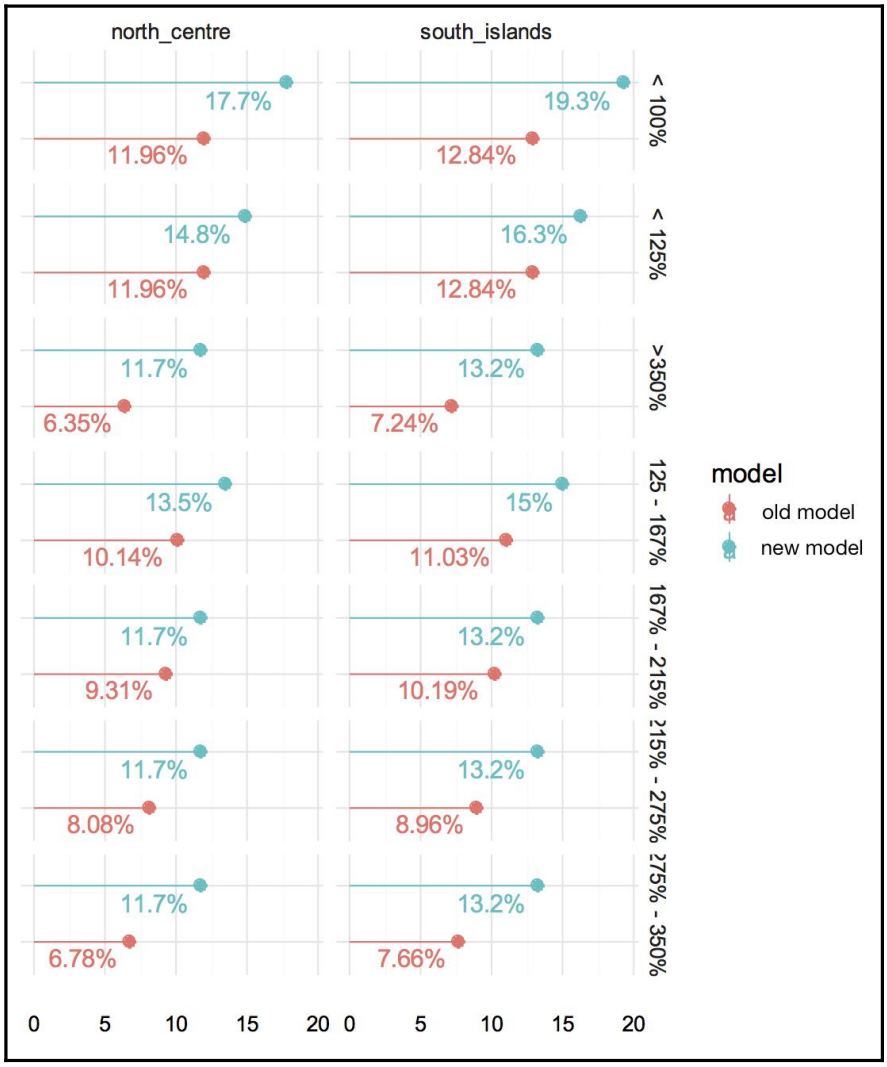
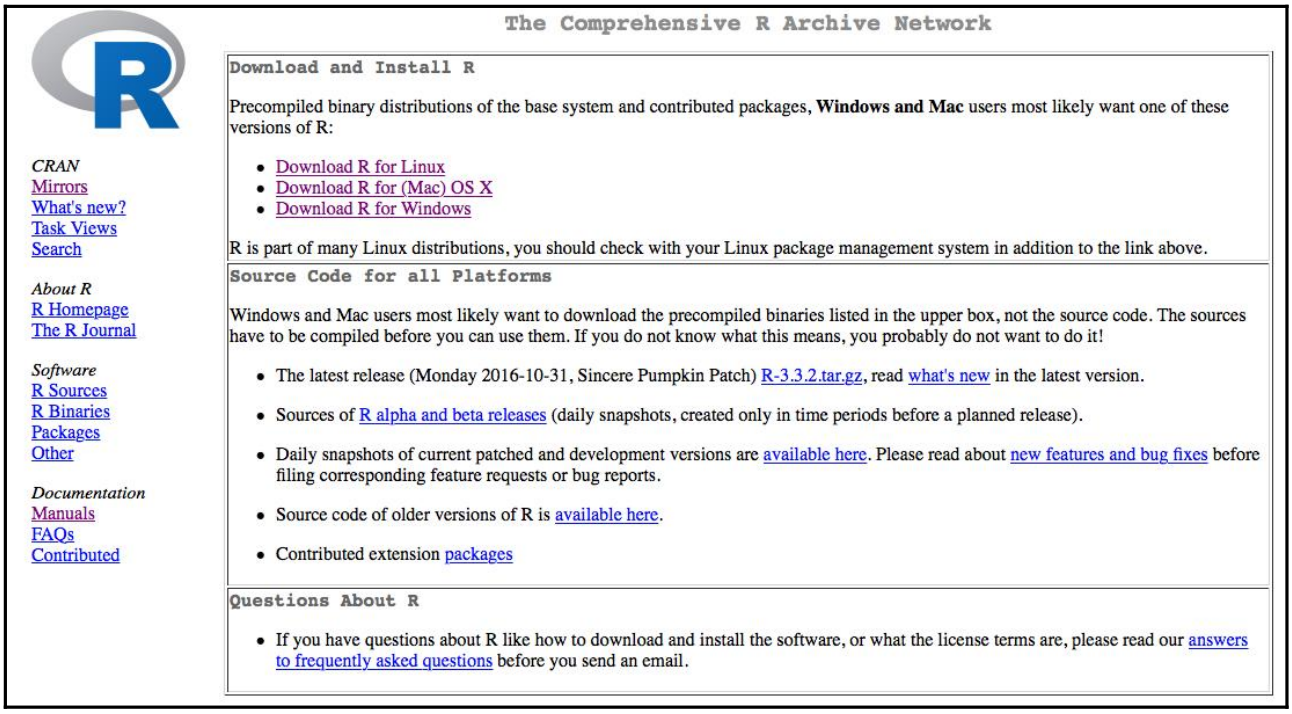
Chapter 1: Why to Choose R for Your Data Mining and Where to Start





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.

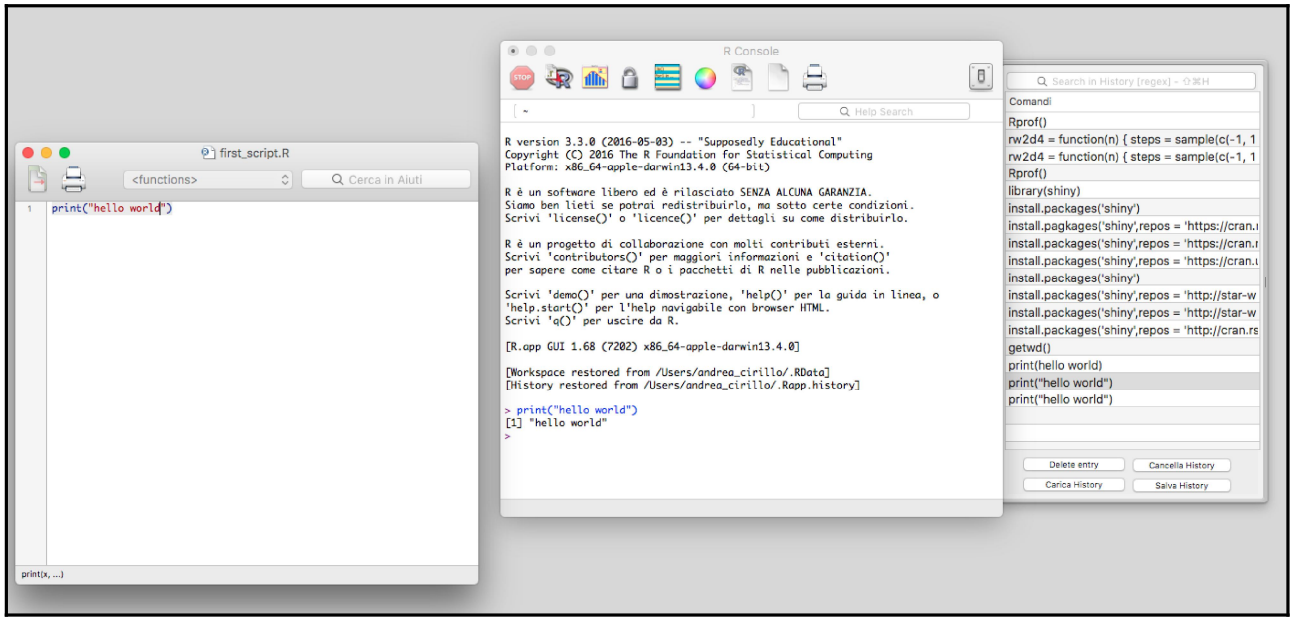
Source Code for all Platforms

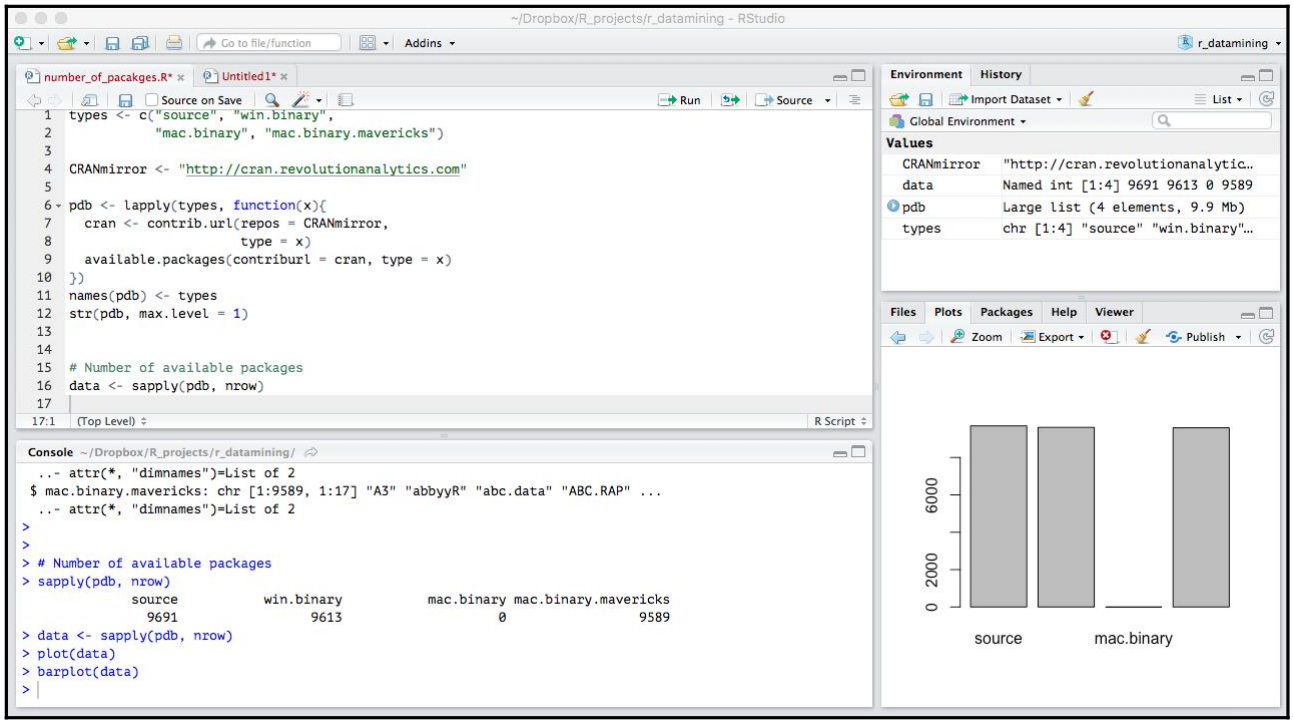
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

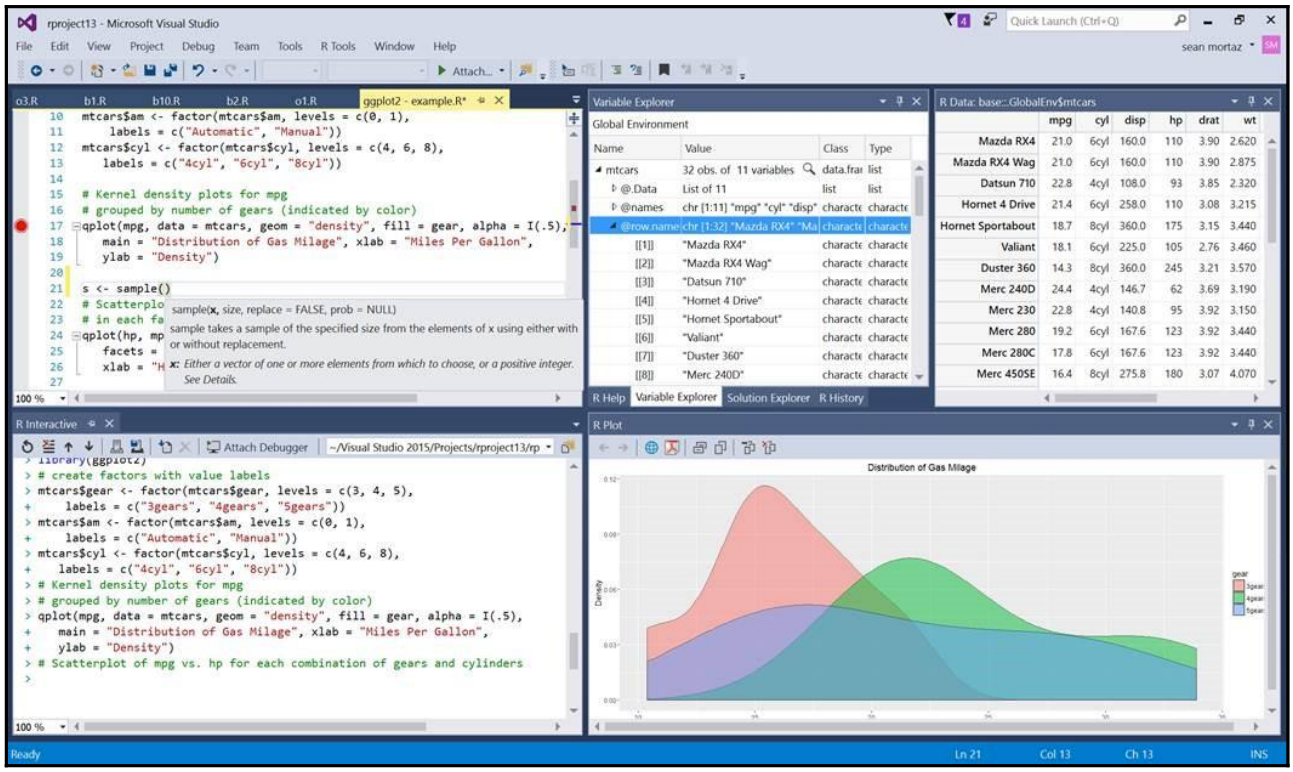
- The latest release (Monday 2016-10-31, Sincere Pumpkin Patch) R-3.3.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available 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

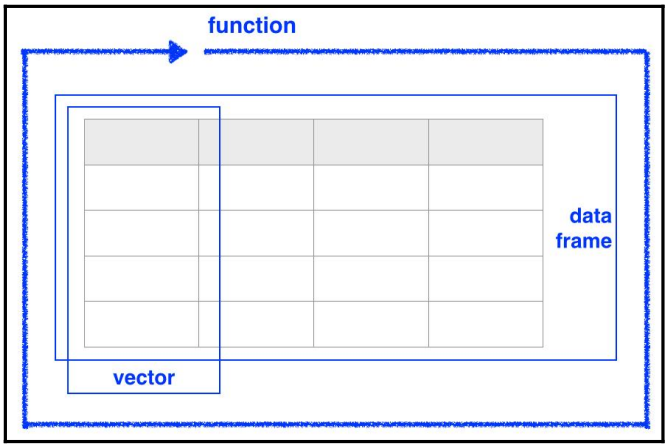
Questions About R

- 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.

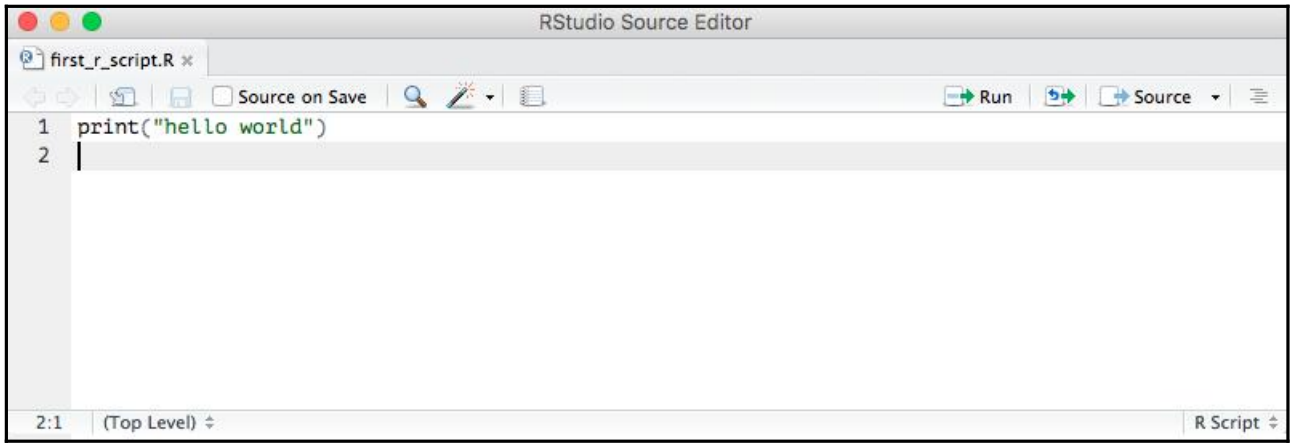


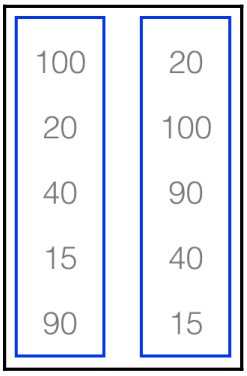

~/Dropbox/R_projects/r_datamining - RStudio
number_of_pacakges.R*
Untitled1*
types <- c("source", "win.binary",
"mac.binary", "mac.binary.mavericks")
CRANmirror <- "http://cran.revolutionanalytics.com"
pdb <- lapply(types, function(x){
cran <- contrib.url(repos = CRANmirror,
type = x)
available.packages(contriburl = cran, type = x)
})
names(pdb) <- types
str(pdb, max.level = 1)
# Number of available packages
data <- sapply(pdb, nrow)
Environment
History
Values
CRANmirror "http://cran.revolutionanalytic…
data Named int [1:4] 9691 9613 0 9589
pdb Large list (4 elements, 9.9 Mb)
types chr [1:4] "source" "win.binary"…
Console ~/Dropbox/R_projects/r_datamining/
> # Number of available packages
> sapply(pdb, nrow)
source win.binary mac.binary mac.binary.mavericks
9691 9613 0 9589
> data <- sapply(pdb, nrow)
> plot(data)
> barplot(data)
Files
Plots
Packages
Help
Viewer
6000
2000
0
source
mac.binary

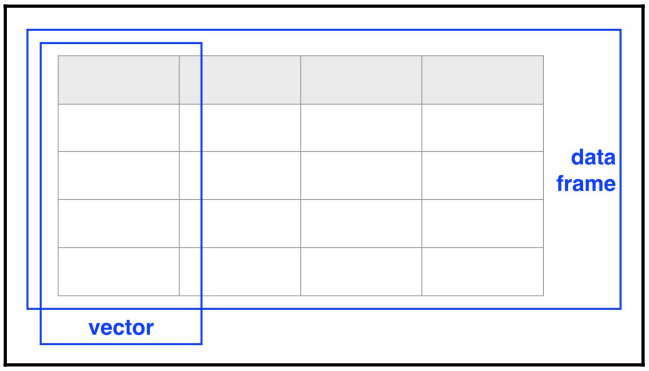

rproject13 - Microsoft Visual Studio
mtcars$am <- factor(mtcars$am, levels = c(0, 1),
labels = c("Automatic", "Manual"))
mtcars$cyl <- factor(mtcars$cyl, levels = c(4, 6, 8),
labels = c("4cyl", "6cyl", "8cyl"))
# Kernel density plots for mpg
# grouped by number of gears (indicated by color)
qplot(mpg, data = mtcars, geom = "density", fill = gear, alpha = I(.5),
main = "Distribution of Gas Milage", xlab = "Miles Per Gallon",
ylab = "Density")
Variable Explorer
Global Environment
R Data: base::GlobalEnv$mtcars
R Interactive
R Plot
Distribution of Gas Milage
Ready

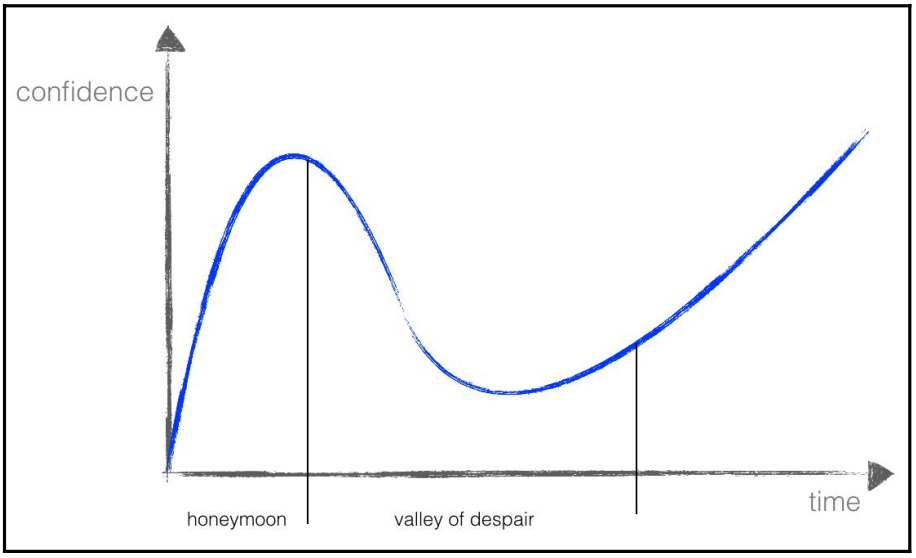



```
Console ~/

> demo()
> demo(scoping)

        demo(scoping)
        ---- ~~~~~~~

Type  <Return>   to start :

> ## Here is a little example which shows a fundamental difference between
> ## R and S.  It is a little example from Abelson and Sussman which models
> ## the way in which bank accounts work.       It shows how R functions can
> ## encapsulate state information.
> ##
> ## When invoked, "open.account" defines and returns three functions
```

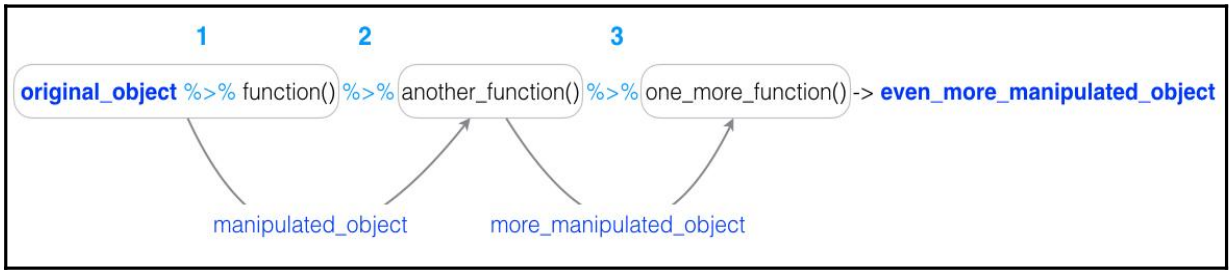


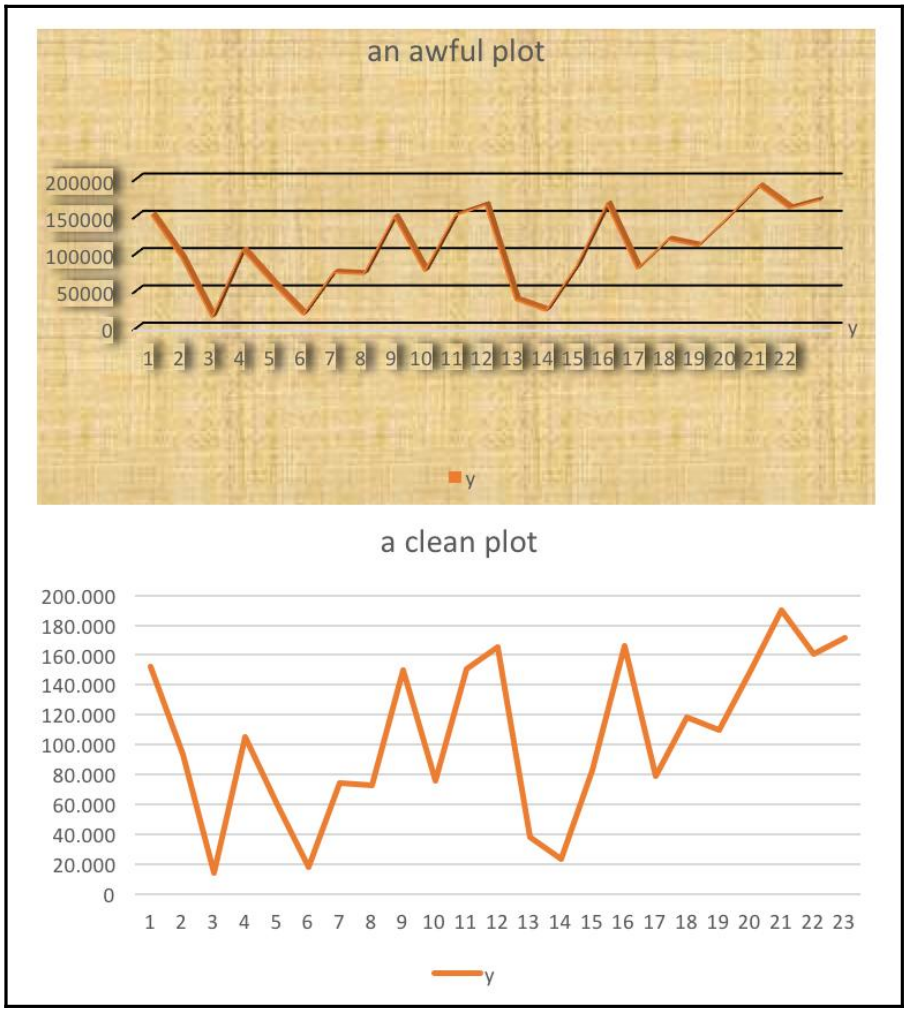

100
20
40
15
90
20
100
90
40
15

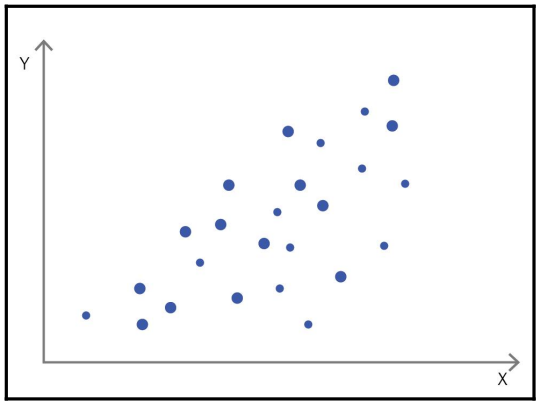

data
frame
vector

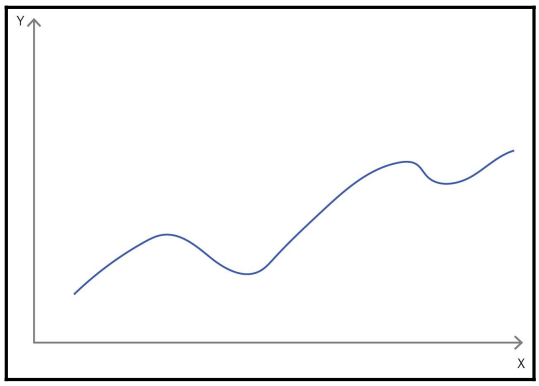

confidence
time
honeymoon
valley of despair

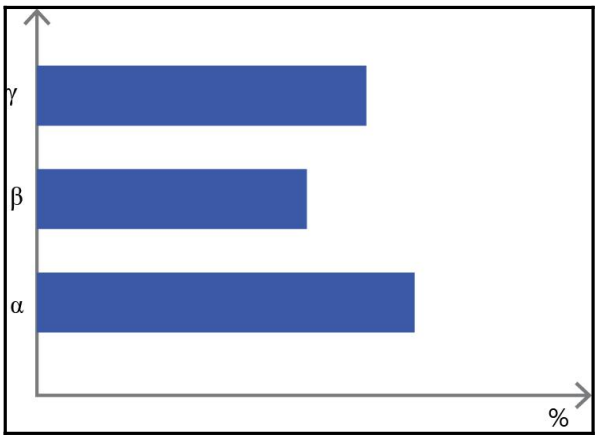
Chapter 2: A First Primer on Data Mining Analysing Your Bank Account Data

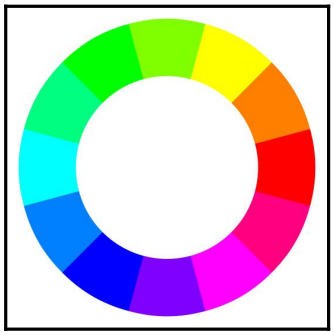


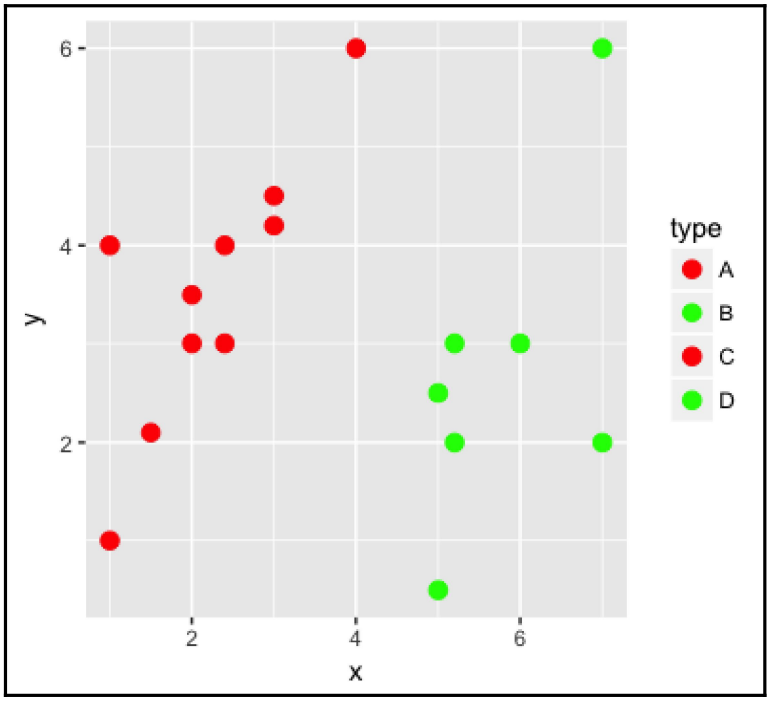

an awful plot
200000
150000
100000
50000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
y
y
a clean plot
200.000
180.000
160.000
140.000
120.000
100.000
80.000
60.000
40.000
20.000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
y

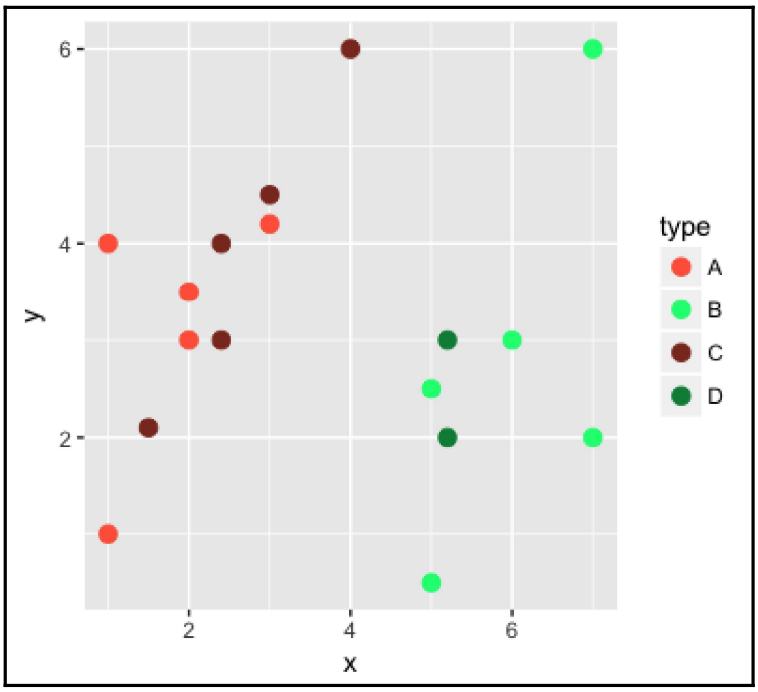

Y
X

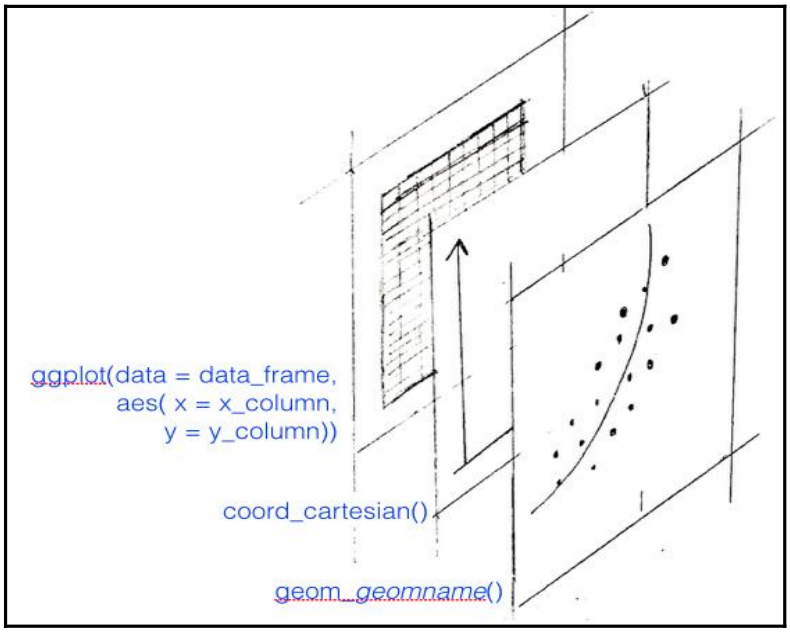

Y
X

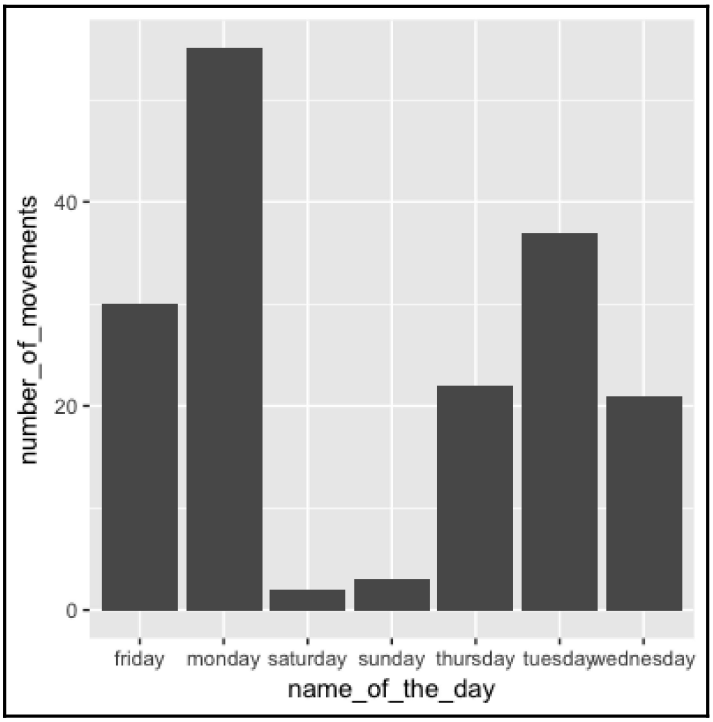

γ
β
α
%

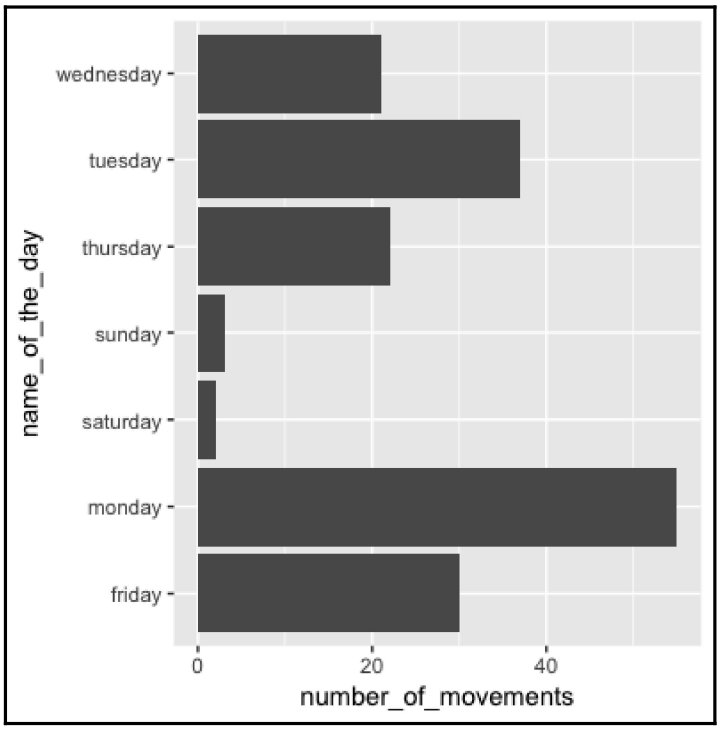


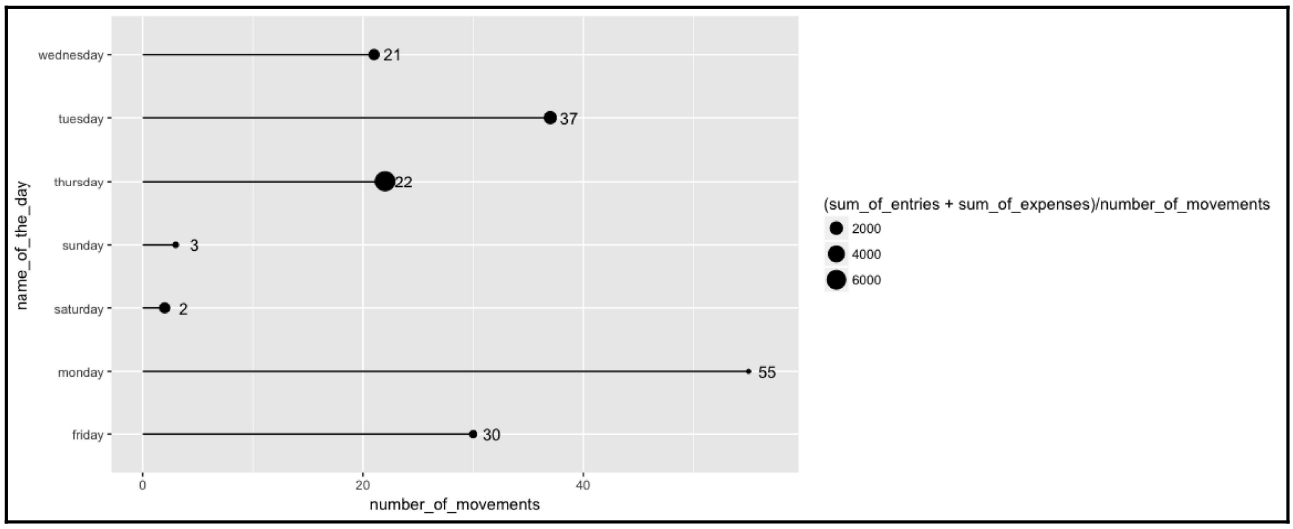

6
4
2
y
2
4
6
x
type
A
B
C
D


6
4
2
y
2
4
6
x
type
A
B
C
D

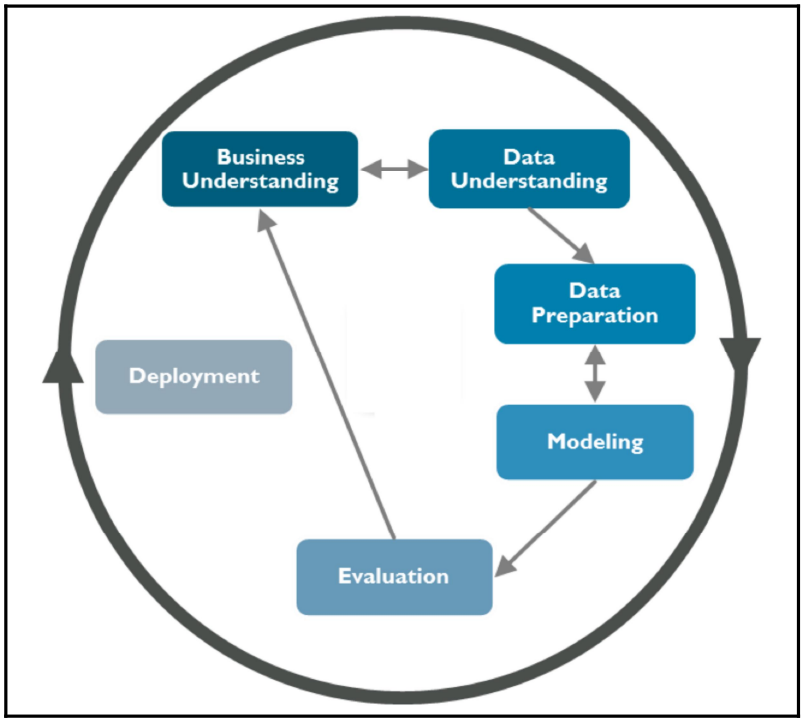

ggplot(data = data_frame,
aes(x = x_column,
y = y_column))
coord_cartesian()
geom_geomname()

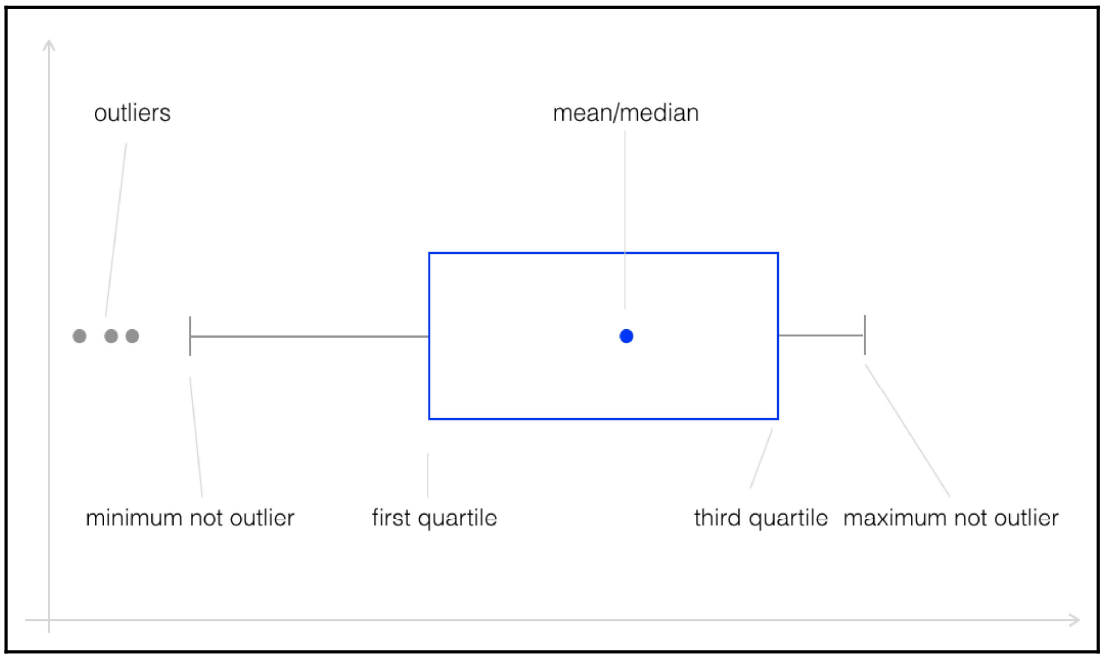

number_of_movements
40
20
0
friday
monday
saturday
sunday
thursday
tuesday
wednesday
name_of_the_day

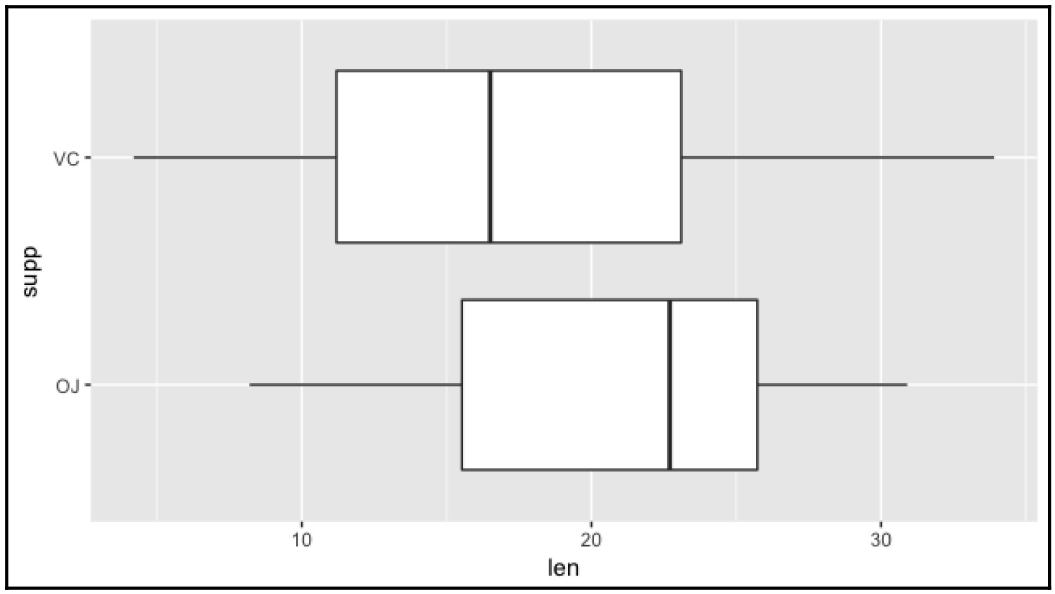

name_of_the_day
wednesday
tuesday
thursday
sunday
saturday
monday
friday
0
20
40
number_of_movements

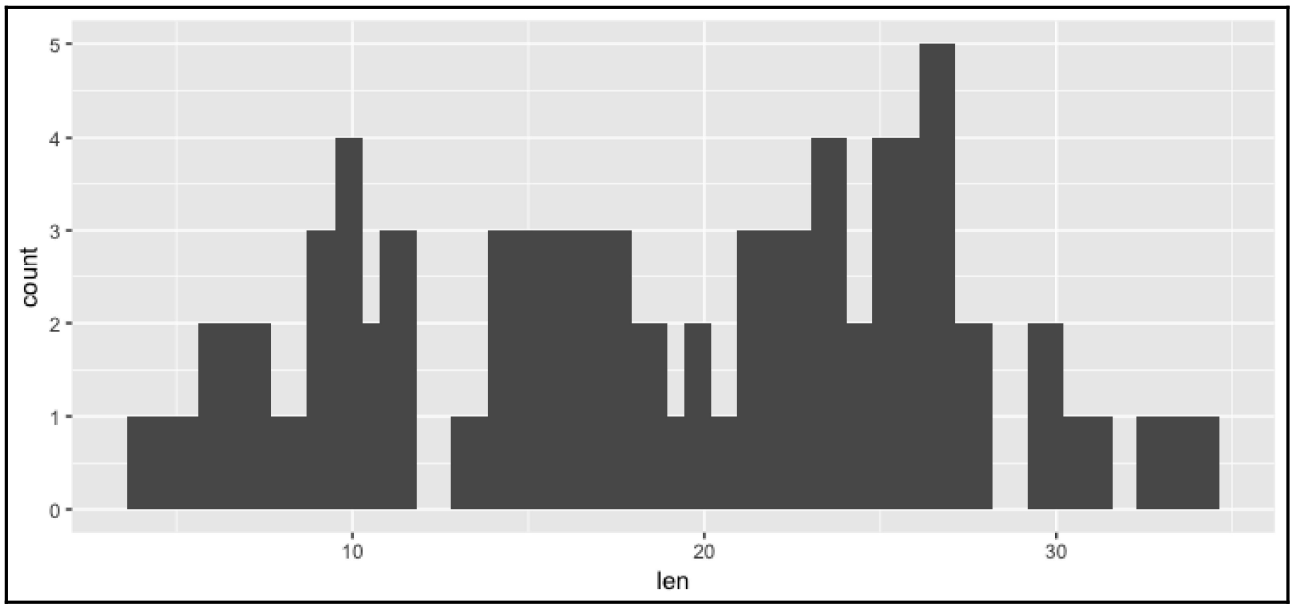

wednesday
tuesday
thursday
sunday
saturday
monday
friday
name_of_the_day
21
37
22
3
2
55
30
0
20
40
number_of_movements
(sum_of_entries + sum_of_expenses)/number_of_movements
2000
4000
6000

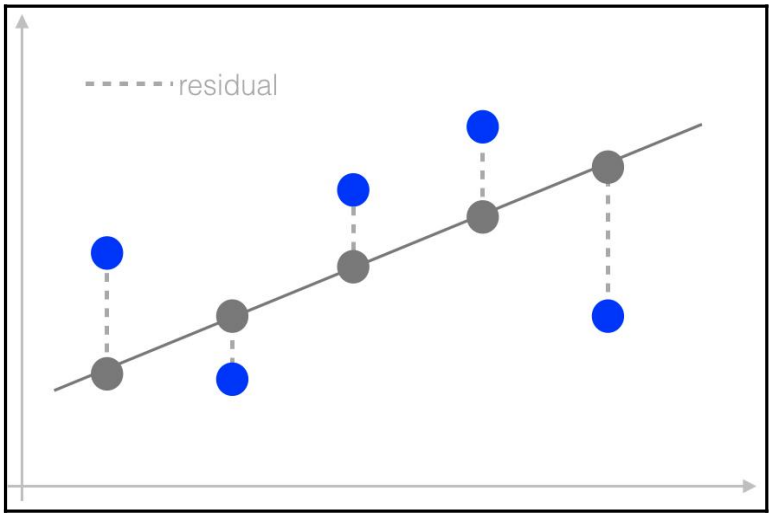
Chapter 3: The Data Mining Process - CRISP-DM Methodology

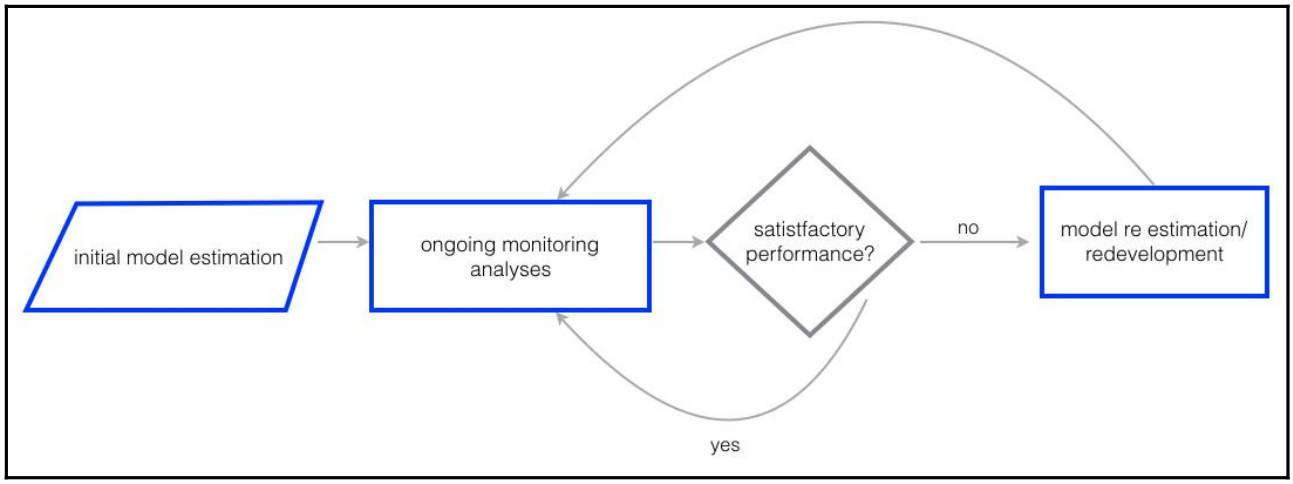



outliers
mean/median
minimum not outlier
first quartile
third quartile
maximum not outlier

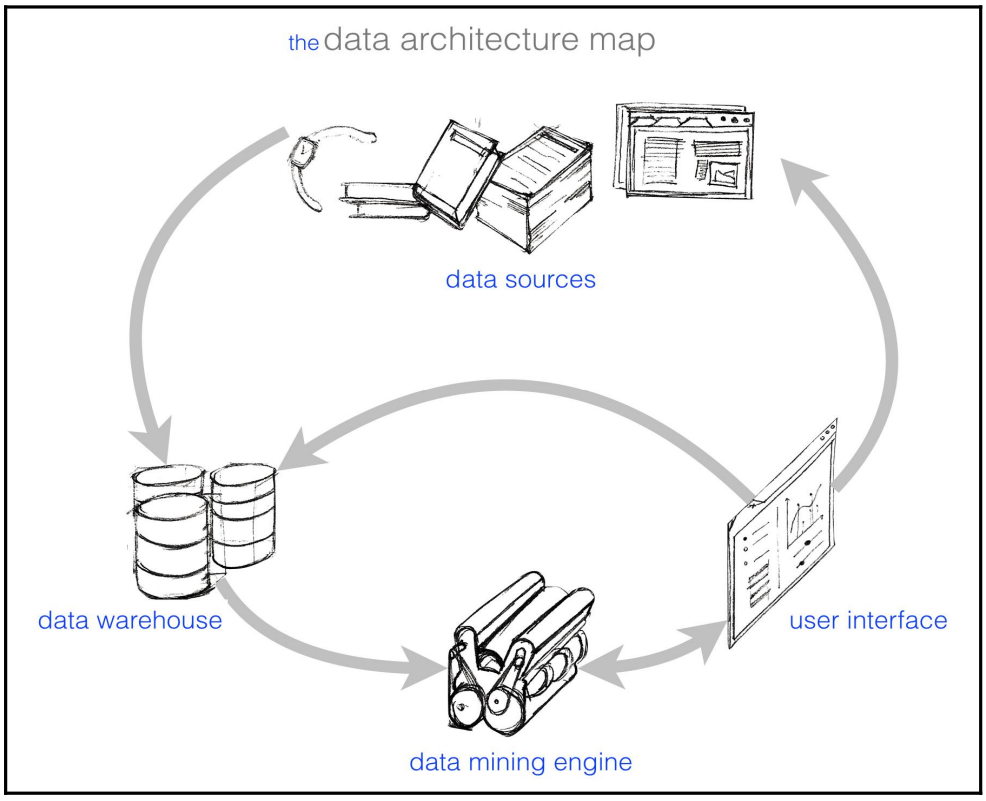

supp
VC
OJ
10
20
30
len

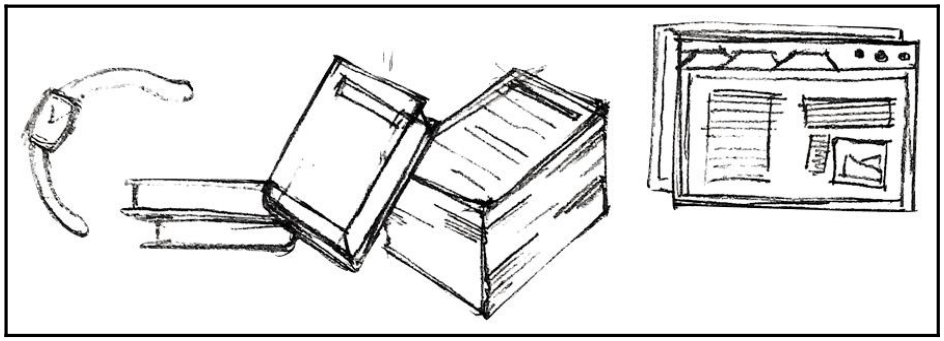

count
len
0
1
2
3
4
5
10
20
30

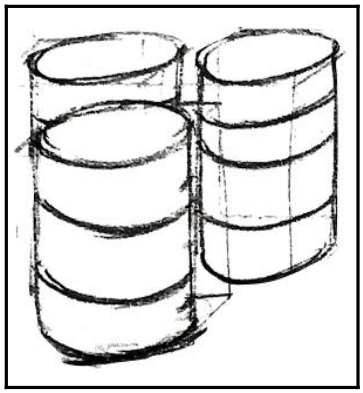

residual

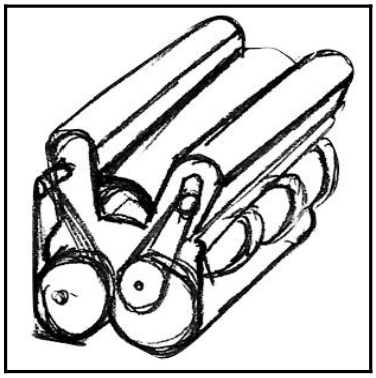

initial model estimation
ongoing monitoring analyses
satistfactory performance?
no
model re estimation/ redevelopment
yes

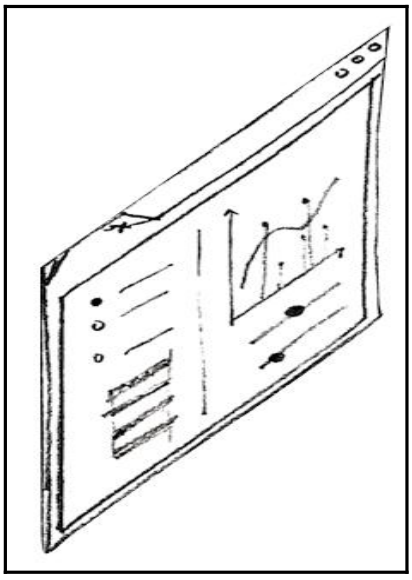
Chapter 4: Keeping the House Clean – The Data Mining Architecture

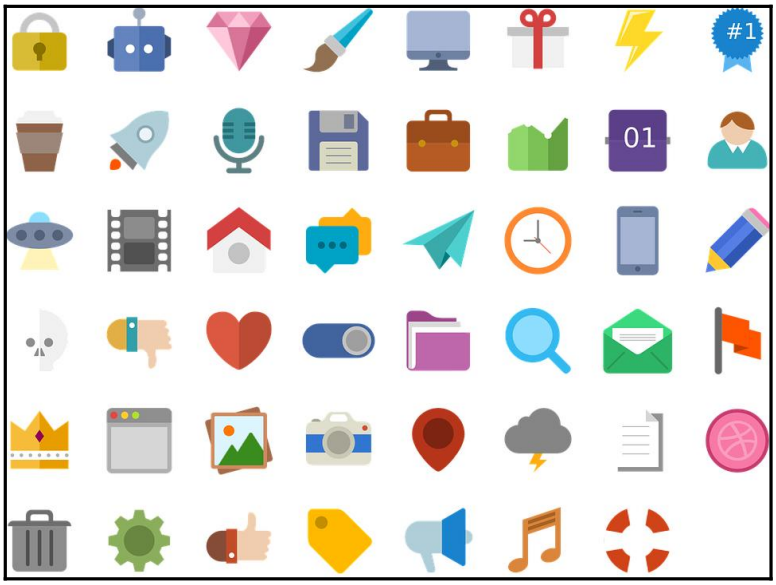




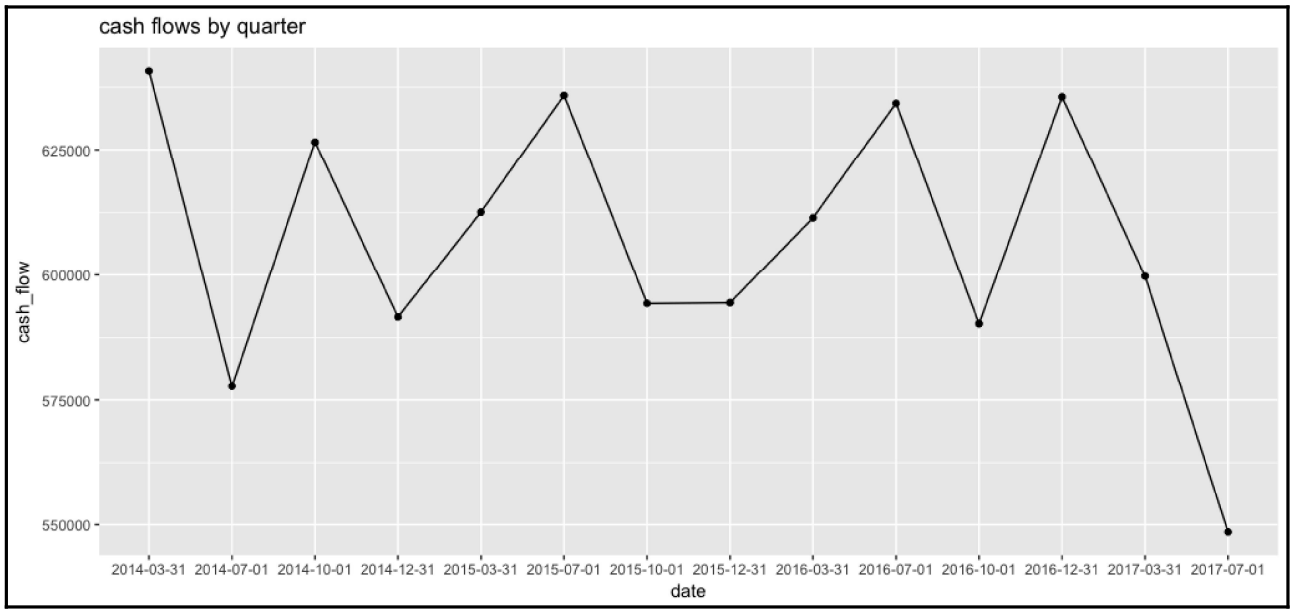




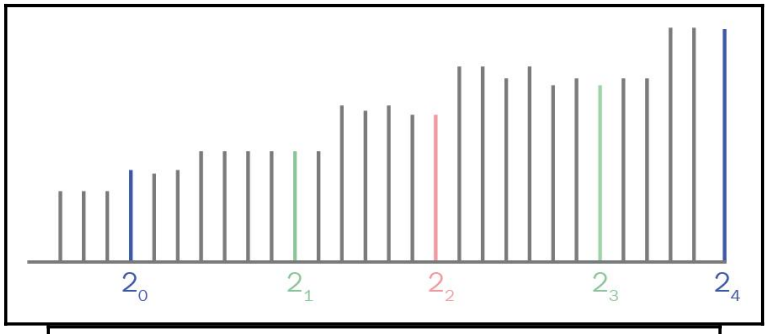


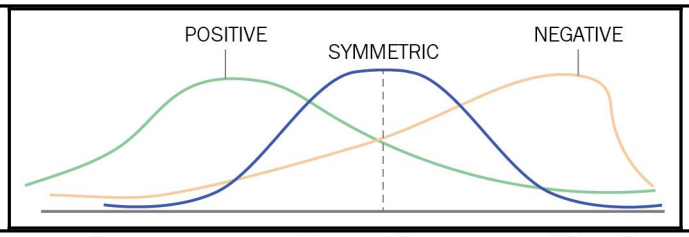

#1
01

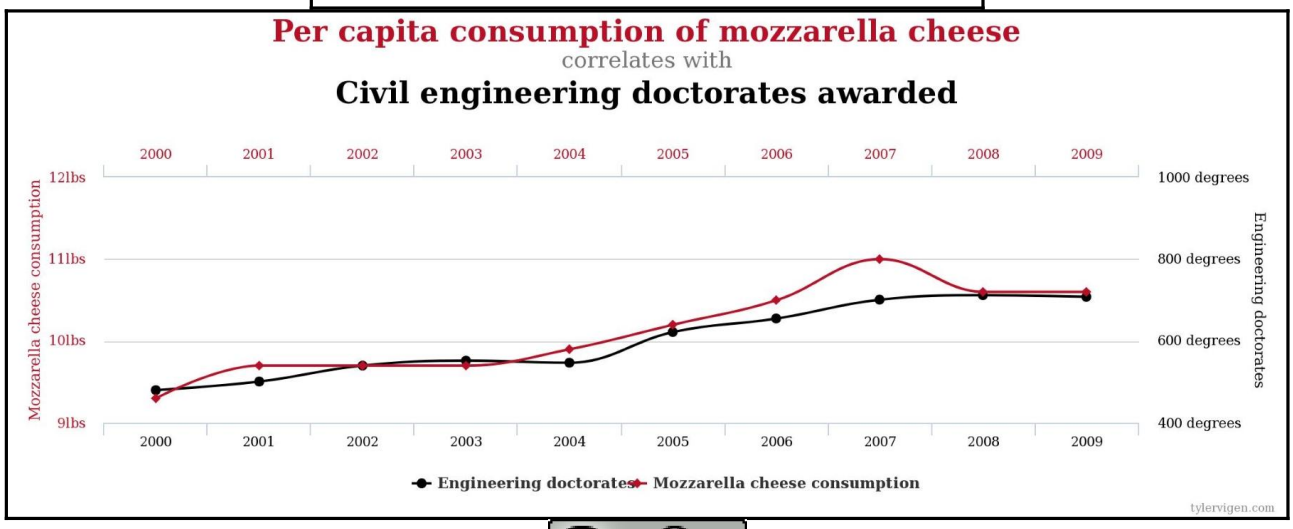
Chapter 5: How to Address a Data Mining Problem – Data Cleaning and Validation

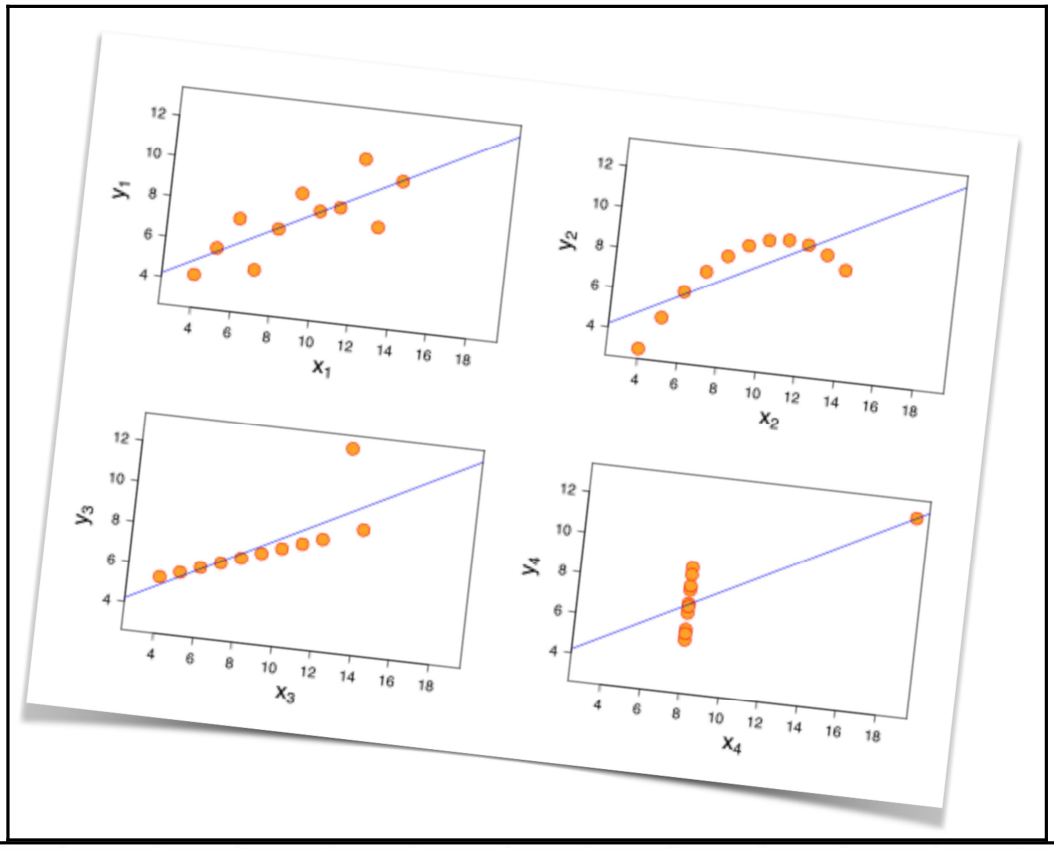


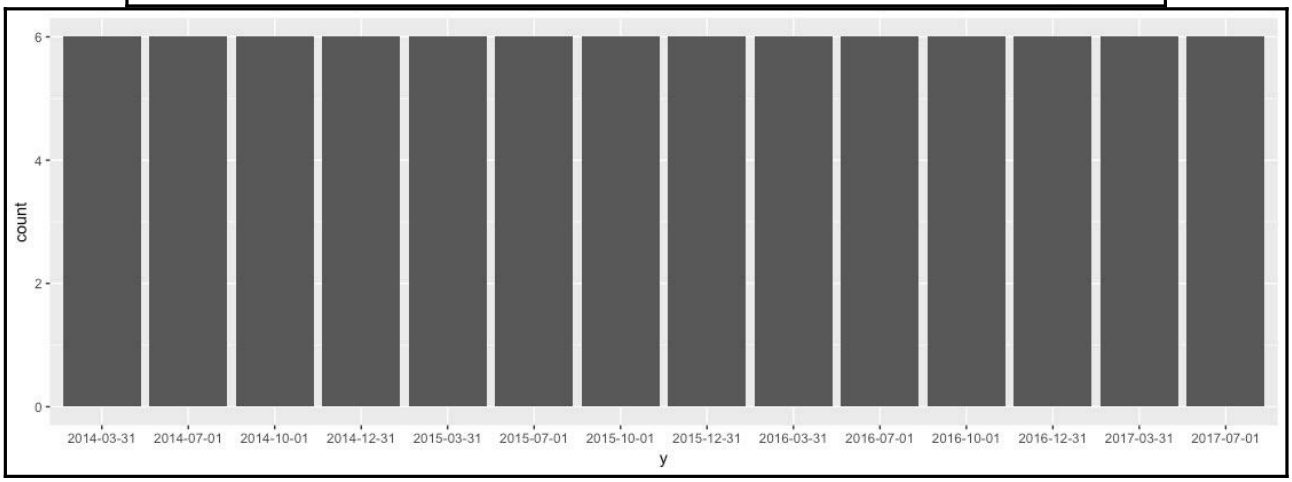
Chapter 6: Looking into Your Data Eyes – Exploratory Data Analysis

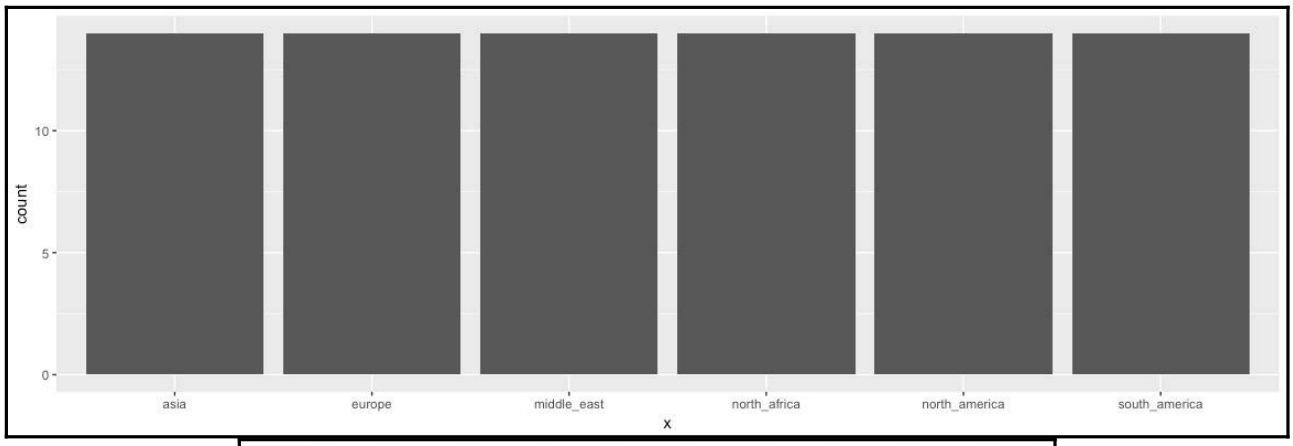


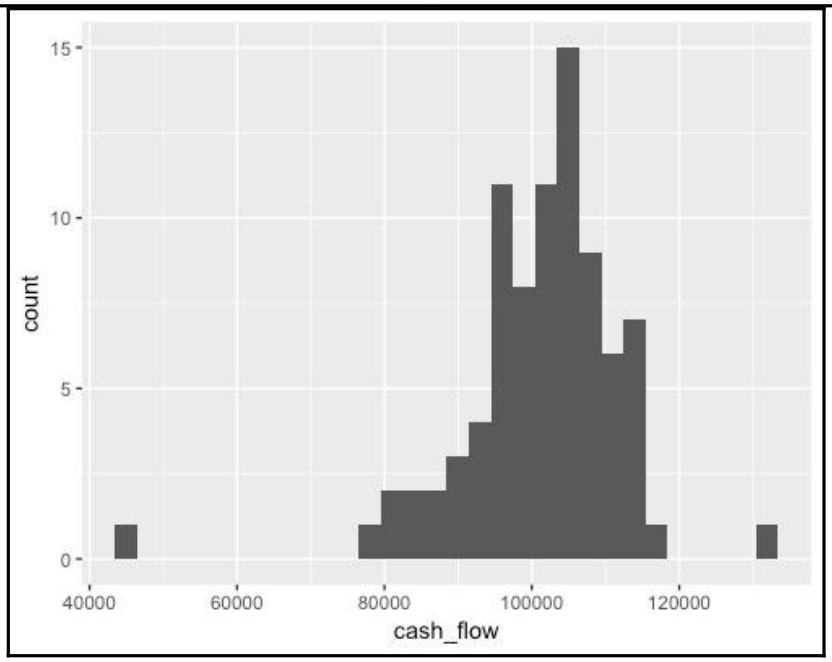


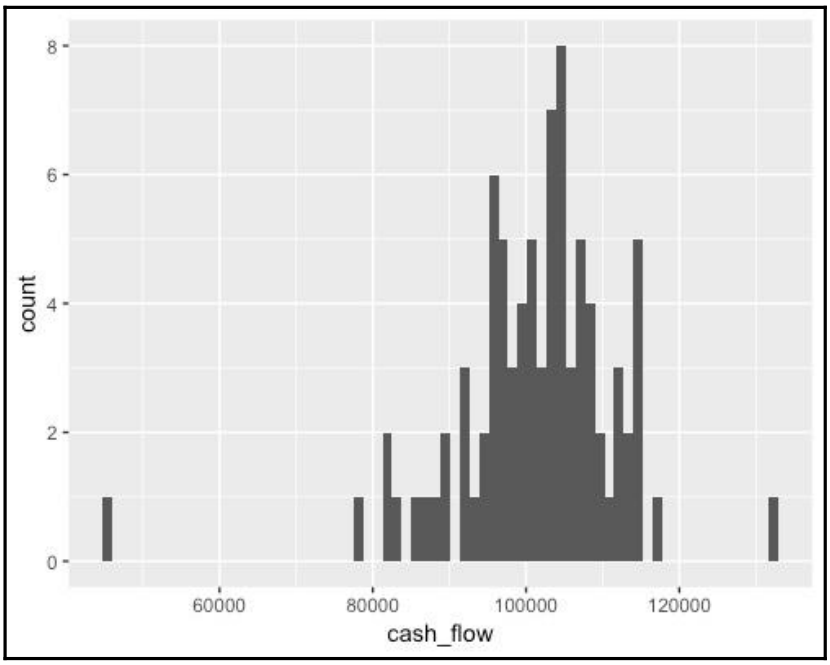


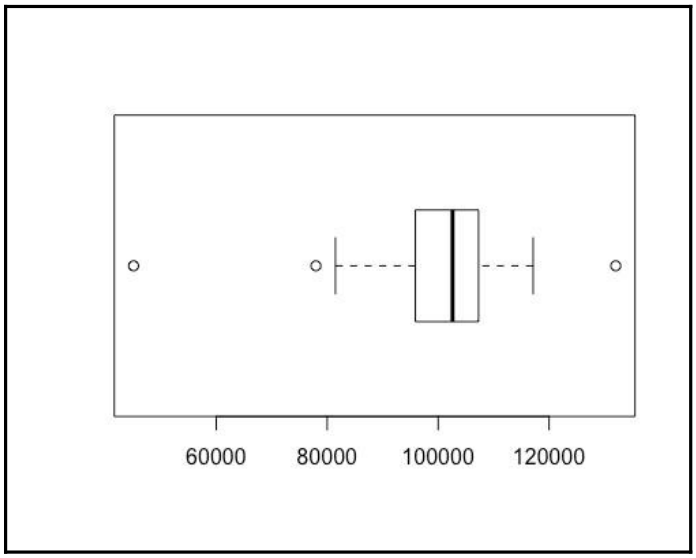

y1
x1
y2
x2
y3
x3
y4
x4

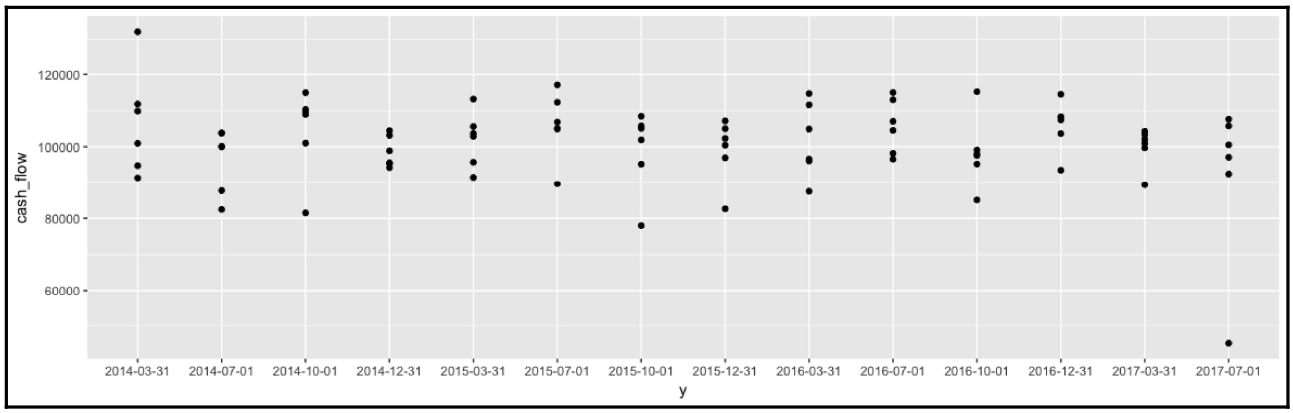

count
y
2014-03-31
2014-07-01
2014-10-01
2014-12-31
2015-03-31
2015-07-01
2015-10-01
2015-12-31
2016-03-31
2016-07-01
2016-10-01
2016-12-31
2017-03-31
2017-07-01

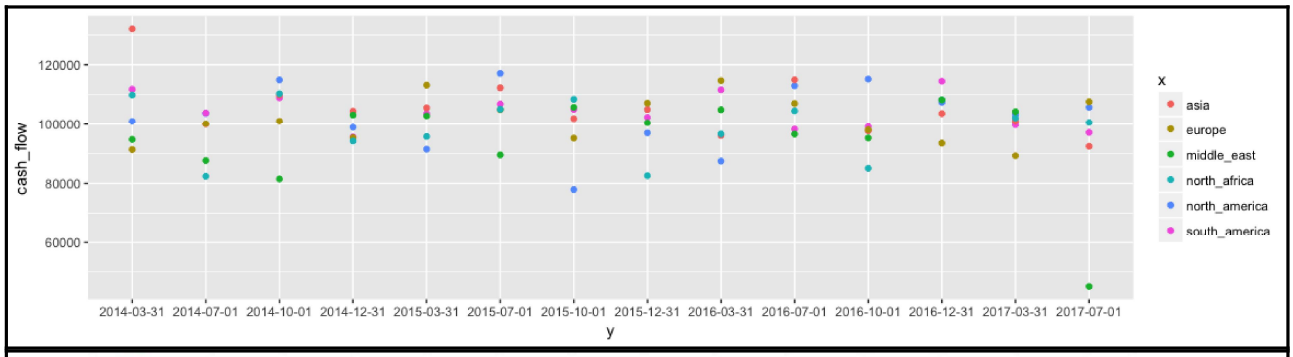

count
10
5
0
asia
europe
middle_east
north_africa
north_america
south_america
x

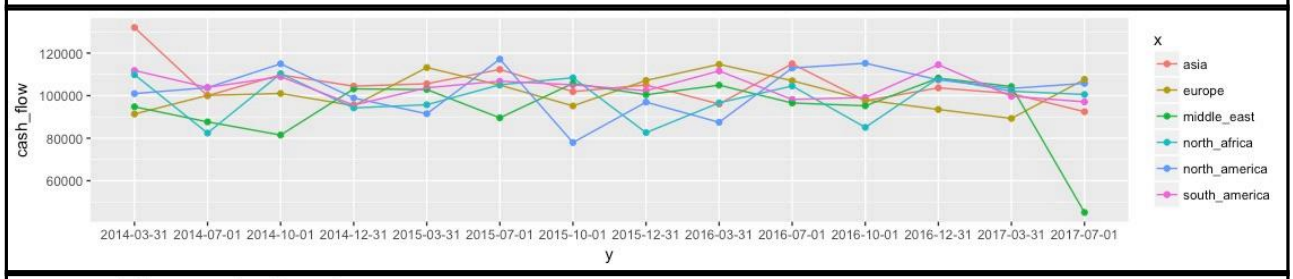

15
10
5
0
count
40000
60000
80000
100000
120000
cash_flow

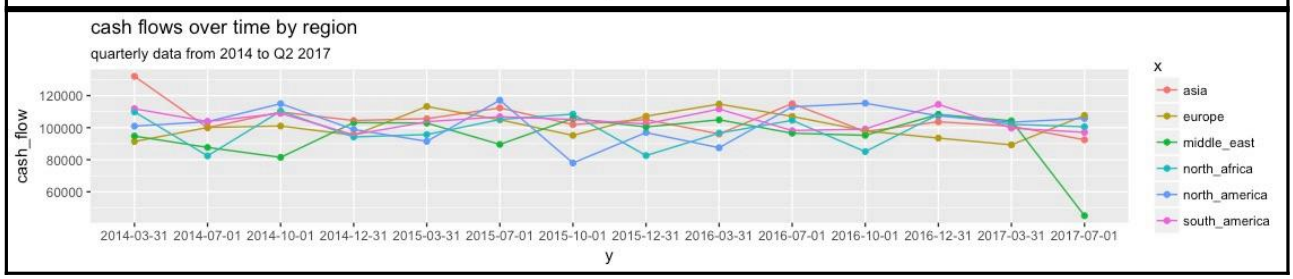

8
6
4
2
0
count
60000
80000
100000
120000
cash_flow

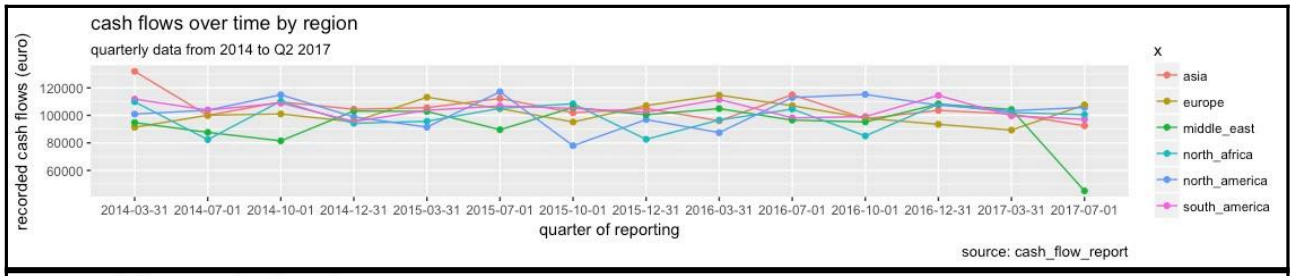

60000
80000
100000
120000

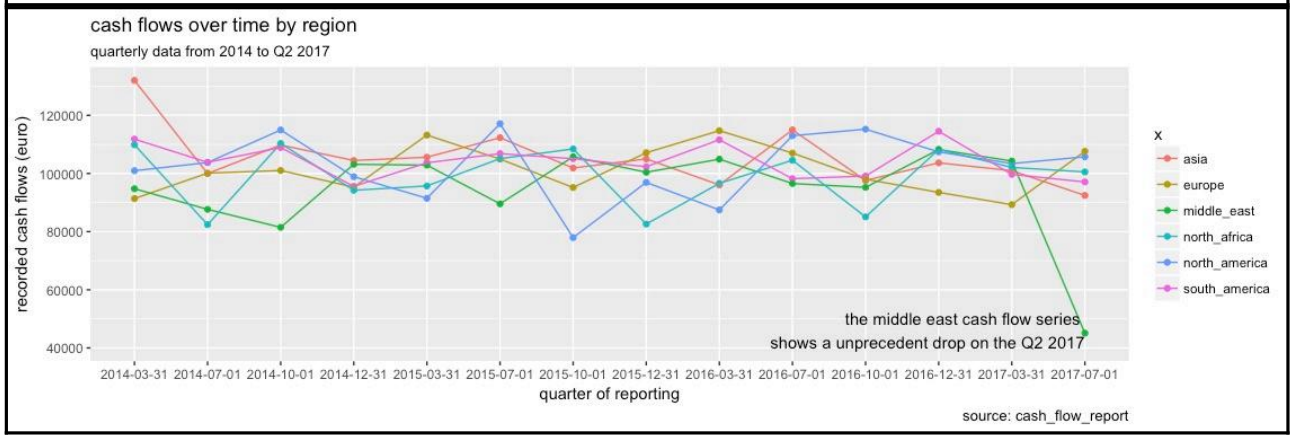

cash_flow
120000
100000
80000
60000
2014-03-31
2014-07-01
2014-10-01
2014-12-31
2015-03-31
2015-07-01
2015-10-01
2015-12-31
2016-03-31
2016-07-01
2016-10-01
2016-12-31
2017-03-31
2017-07-01
y

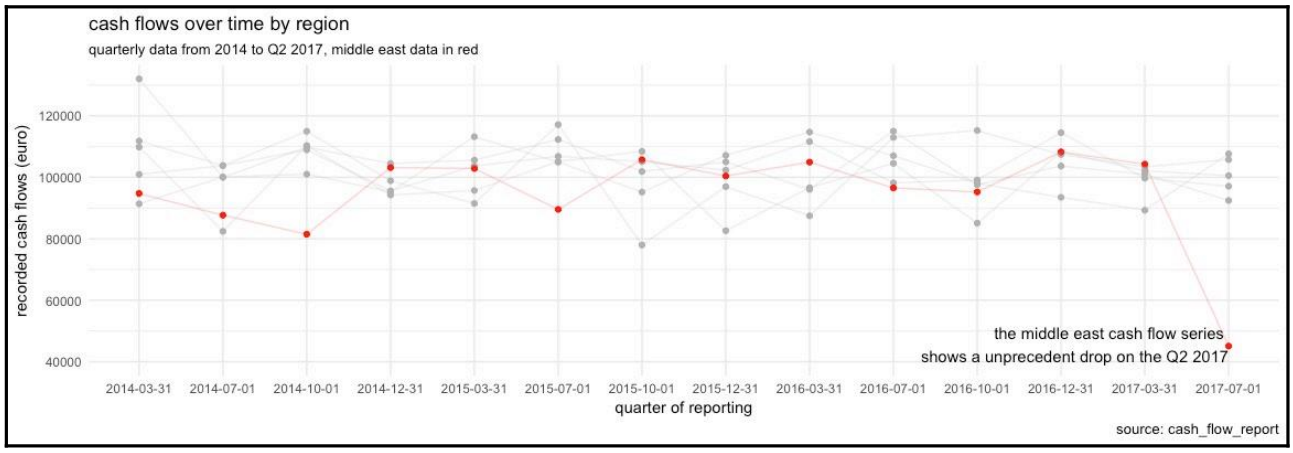

cash_flow
120000
100000
80000
60000
2014-03-31
2014-07-01
2014-10-01
2014-12-31
2015-03-31
2015-07-01
2015-10-01
2015-12-31
2016-03-31
2016-07-01
2016-10-01
2016-12-31
2017-03-31
2017-07-01
y
x
asia
europe
middle_east
north_africa
north_america
south_america


cash_flow
120000
100000
80000
60000
2014-03-31
2014-07-01
2014-10-01
2014-12-31
2015-03-31
2015-07-01
2015-10-01
2015-12-31
2016-03-31
2016-07-01
2016-10-01
2016-12-31
2017-03-31
2017-07-01
y
x
asia
europe
middle_east
north_africa
north_america
south_america

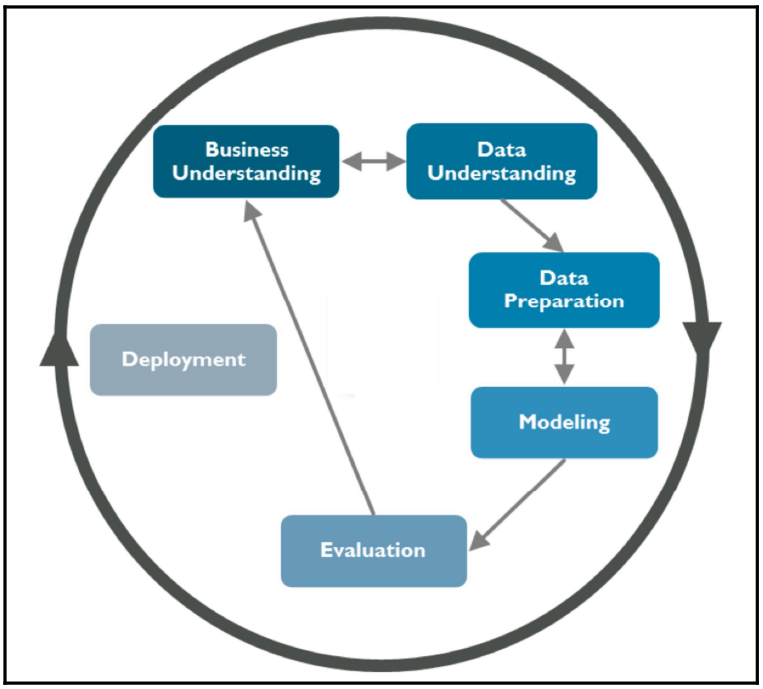

cash flows over time by region
quarterly data from 2014 to Q2 2017
cash_flow
120000
100000
80000
60000
2014-03-31
2014-07-01
2014-10-01
2014-12-31
2015-03-31
2015-07-01
2015-10-01
2015-12-31
2016-03-31
2016-07-01
2016-10-01
2016-12-31
2017-03-31
2017-07-01
y
x
asia
europe
middle_east
north_africa
north_america
south_america

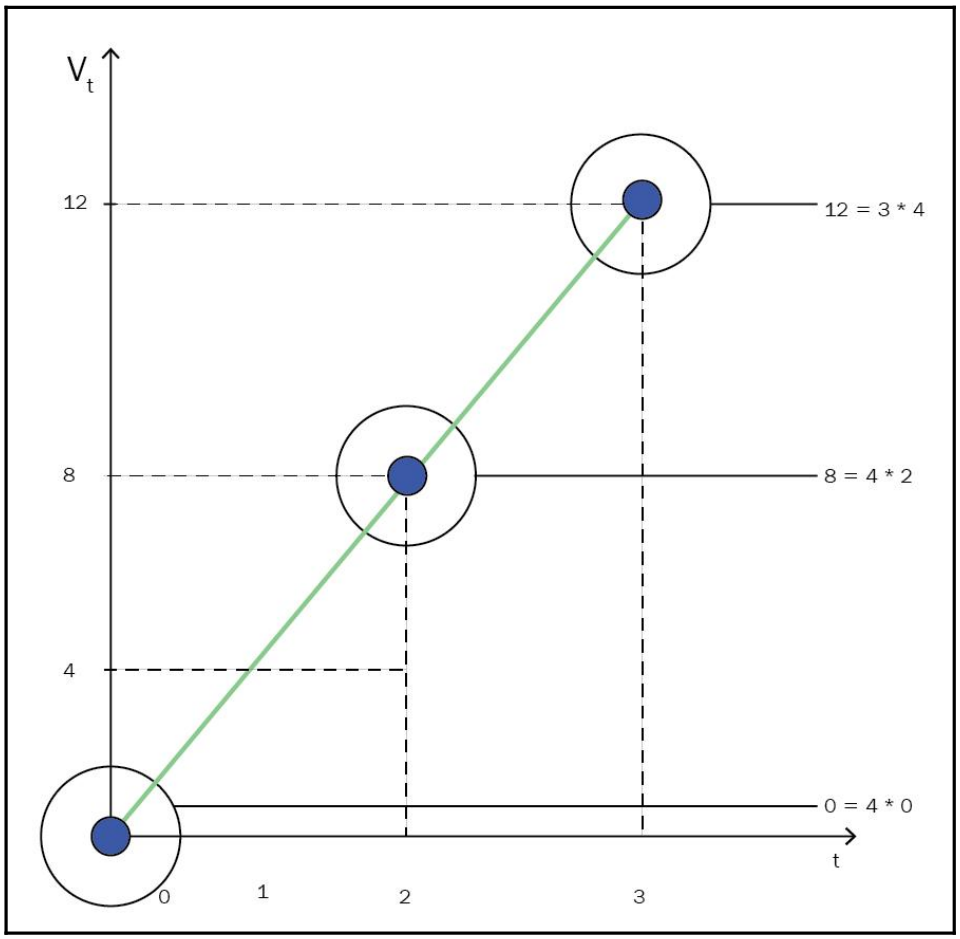

cash flows over time by region
quarterly data from 2014 to Q2 2017
recorded cash flows (euro)
120000
100000
80000
60000
2014-03-31 2014-07-01 2014-10-01 2014-12-31 2015-03-31 2015-07-01 2015-10-01 2015-12-31 2016-03-31 2016-07-01 2016-10-01 2016-12-31 2017-03-31 2017-07-01
quarter of reporting
source: cash_flow_report
x
asia
europe
middle_east
north_africa
north_america
south_america

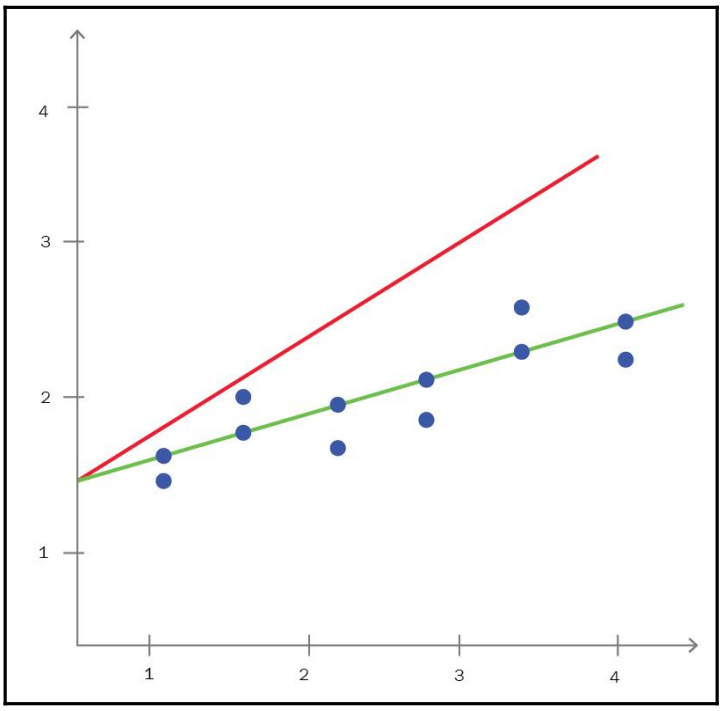

cash flows over time by region
quarterly data from 2014 to Q2 2017
recorded cash flows (euro)
120000
100000
80000
60000
40000
the middle east cash flow series
shows a unprecedent drop on the Q2 2017
2014-03-31 2014-07-01 2014-10-01 2014-12-31 2015-03-31 2015-07-01 2015-10-01 2015-12-31 2016-03-31 2016-07-01 2016-10-01 2016-12-31 2017-03-31 2017-07-01
quarter of reporting
source: cash_flow_report
x
asia
europe
middle_east
north_africa
north_america
south_america

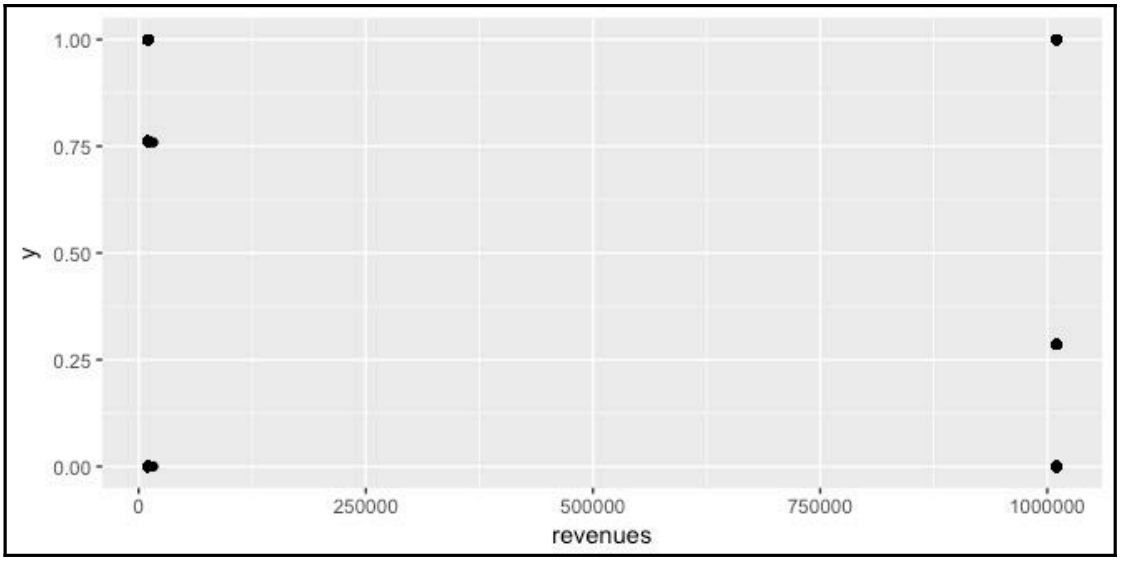

cash flows over time by region
quarterly data from 2014 to Q2 2017, middle east data in red
recorded cash flows (euro)
120000
100000
80000
60000
40000
the middle east cash flow series
shows a unprecedent drop on the Q2 2017
2014-03-31 2014-07-01 2014-10-01 2014-12-31 2015-03-31 2015-07-01 2015-10-01 2015-12-31 2016-03-31 2016-07-01 2016-10-01 2016-12-31 2017-03-31 2017-07-01
quarter of reporting
source: cash_flow_report

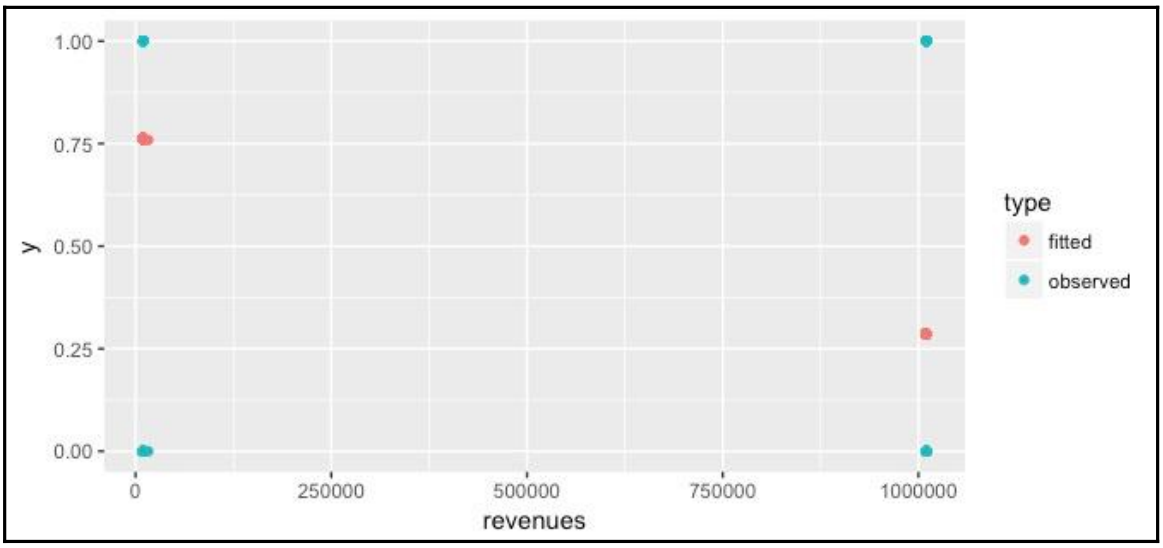
Chapter 7: Our First Guess – a Linear Regression



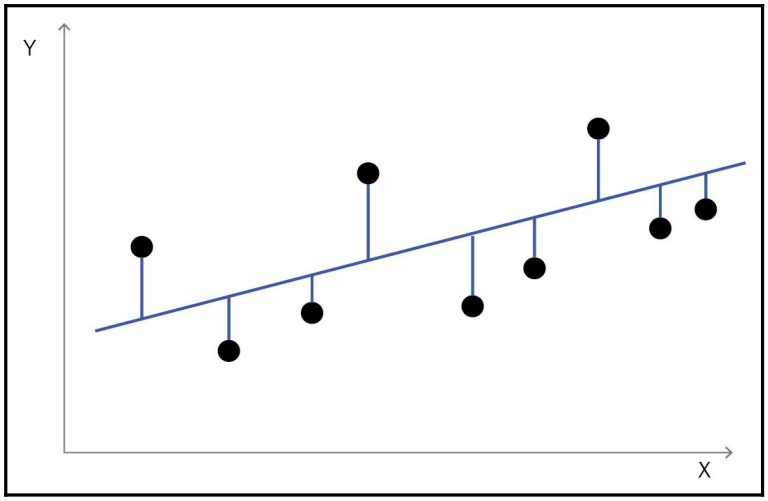

V_t
12
8
4
12 = 3 * 4
8 = 4 * 2
0 = 4 * 0
t
0
1
2
3

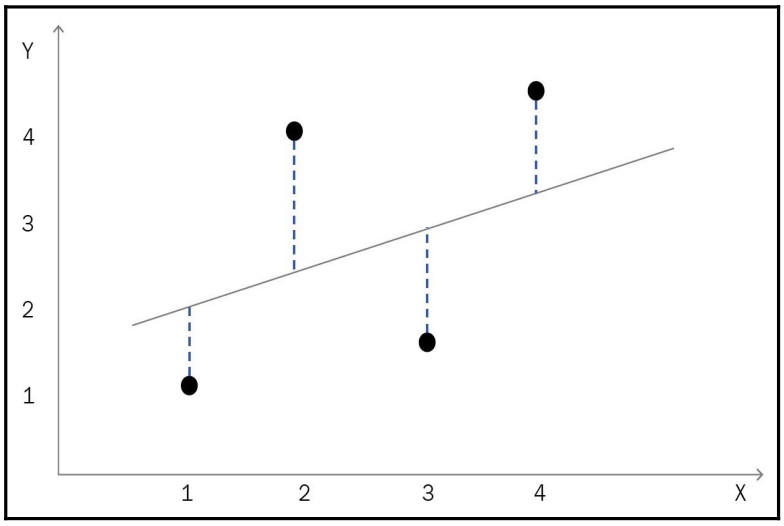

4
3
2
1
1
2
3
4

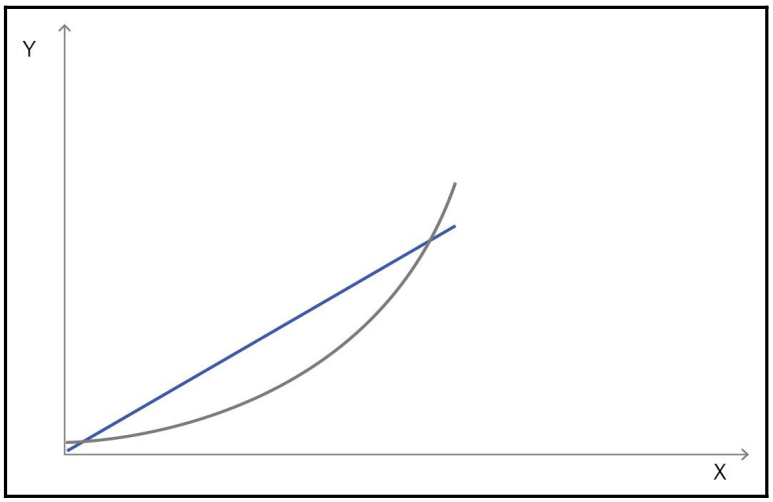

1.00
0.75
0.50
0.25
0.00
y
0
250000
500000
750000
1000000
revenues

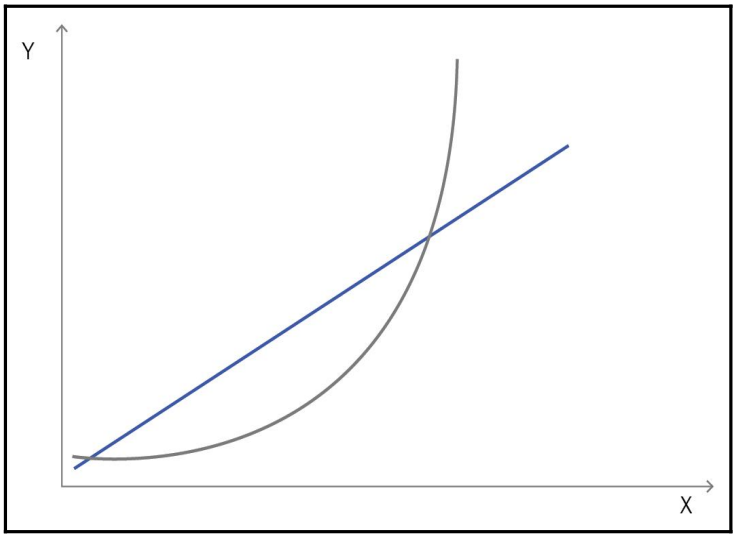

1.00
0.75
0.50
0.25
0.00
y
0
250000
500000
750000
1000000
revenues
type
fitted
observed

Chapter 8: A Gentle Introduction to Model Performance Evaluation

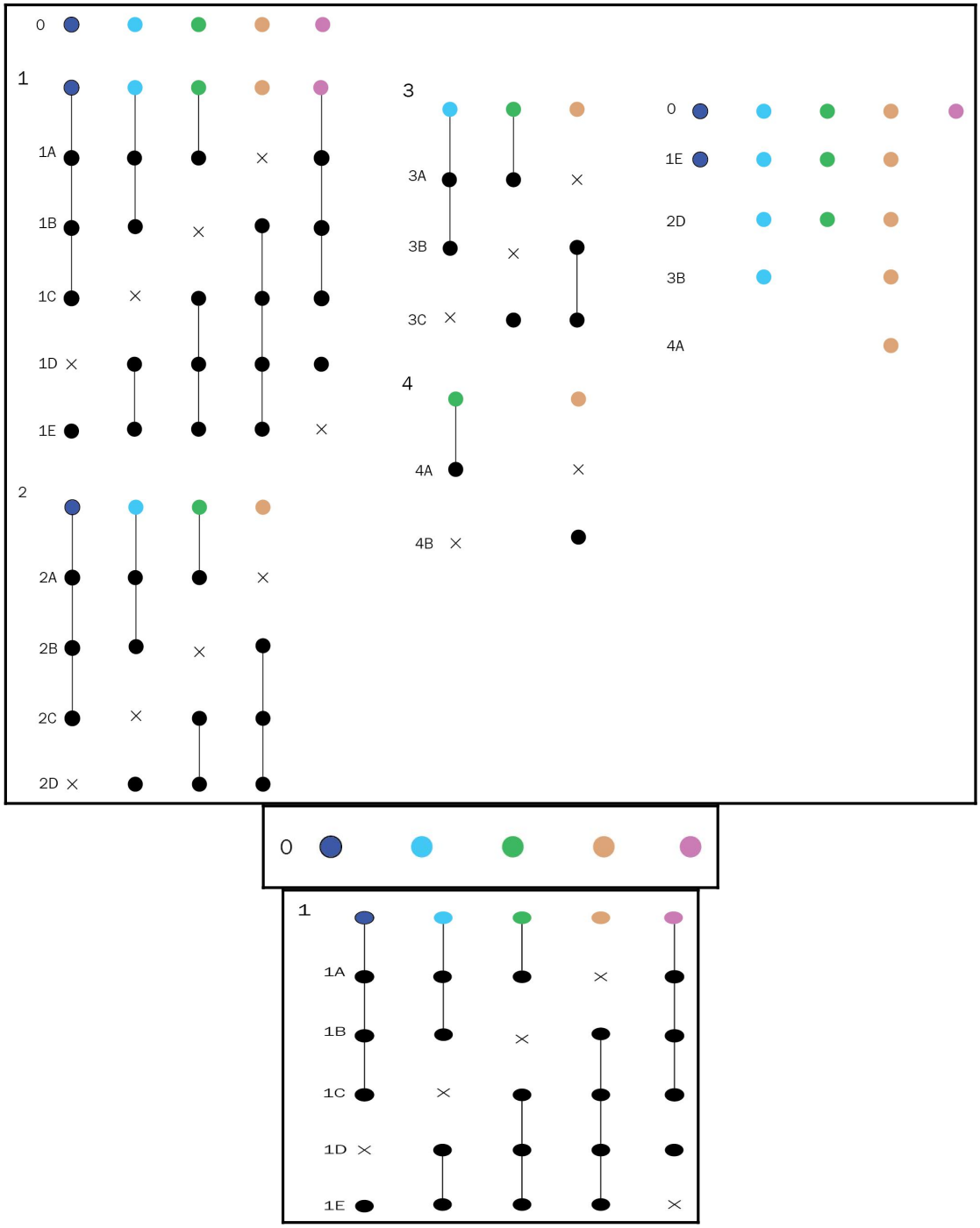


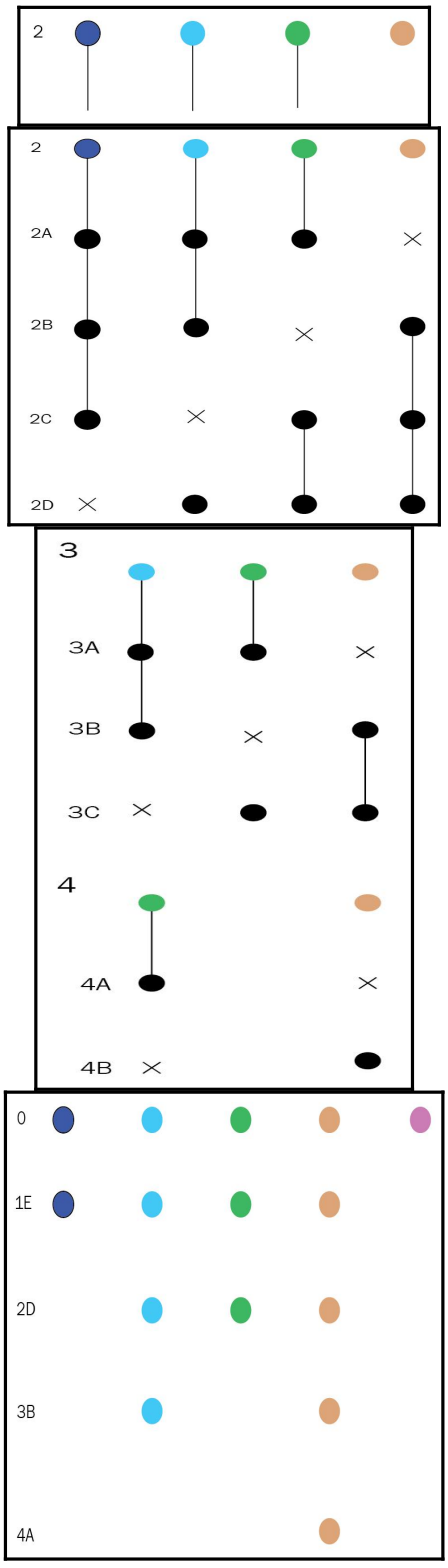


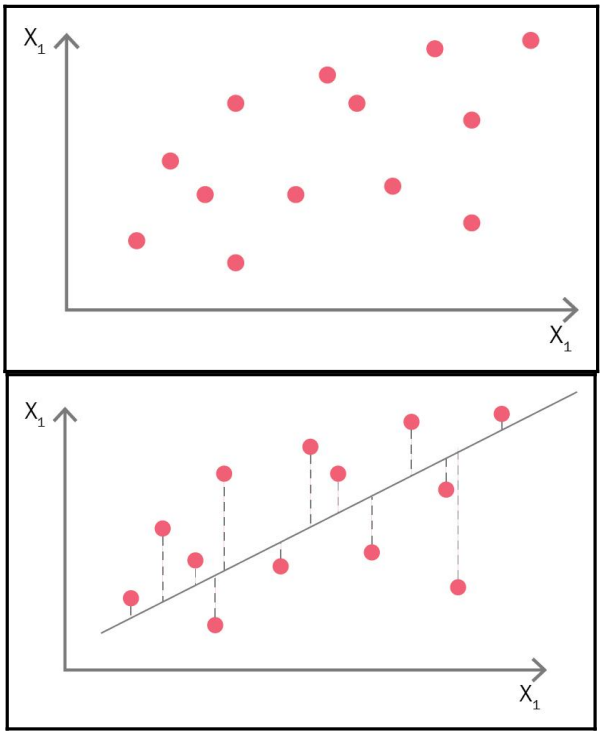

Y
X

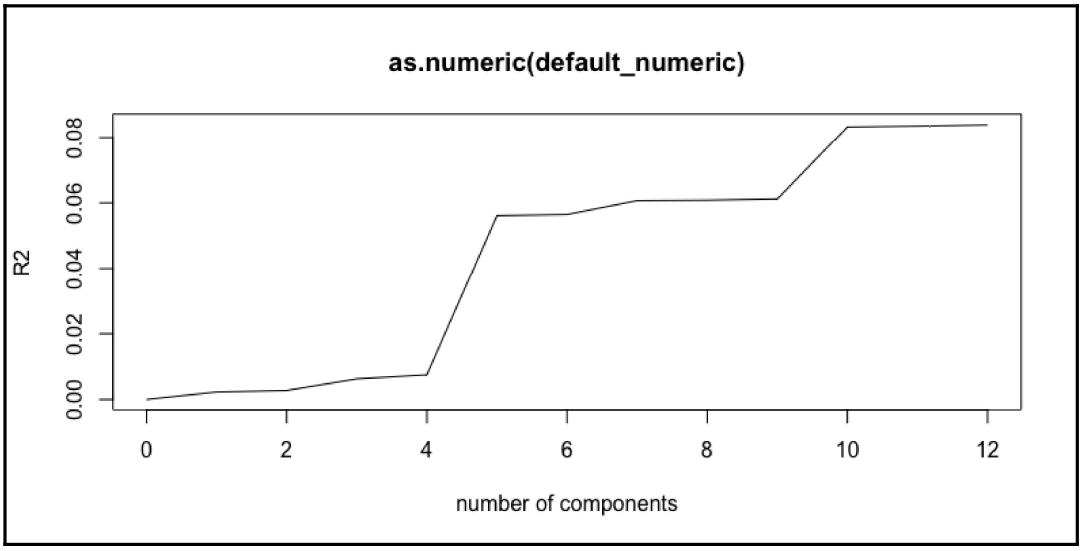

Y
X

Chapter 9: Don't Give up – Power up Your Regression Including Multiple Variables

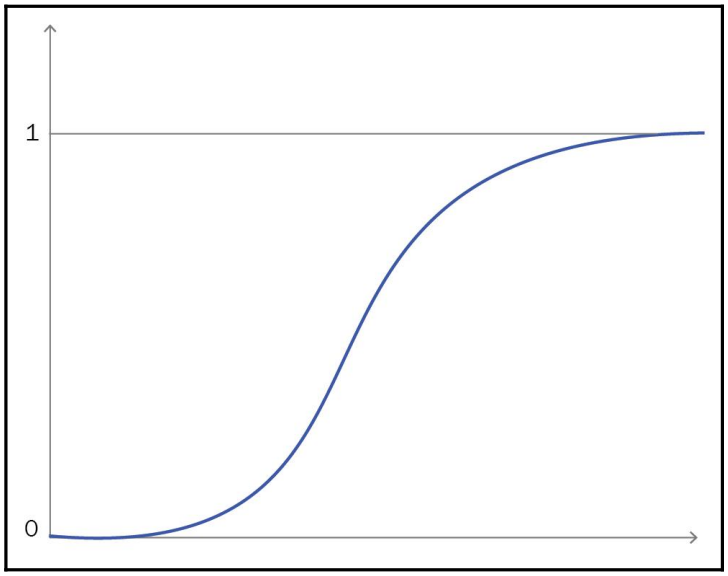


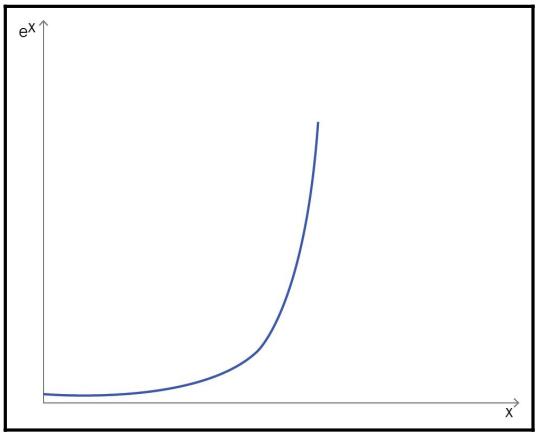

2
2
2A
2B
2C
2D
3
3A
3B
3C
4
4A
4B
0
1E
2D
3B
4A

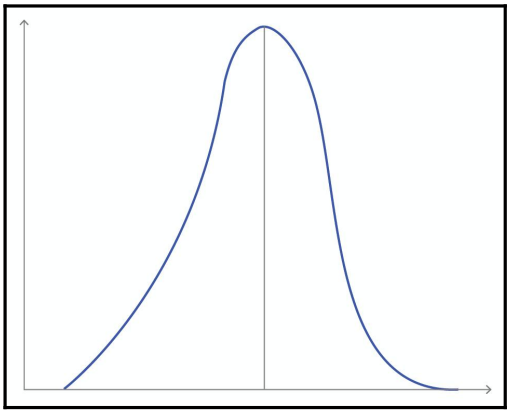

X1
X1
X1
X1

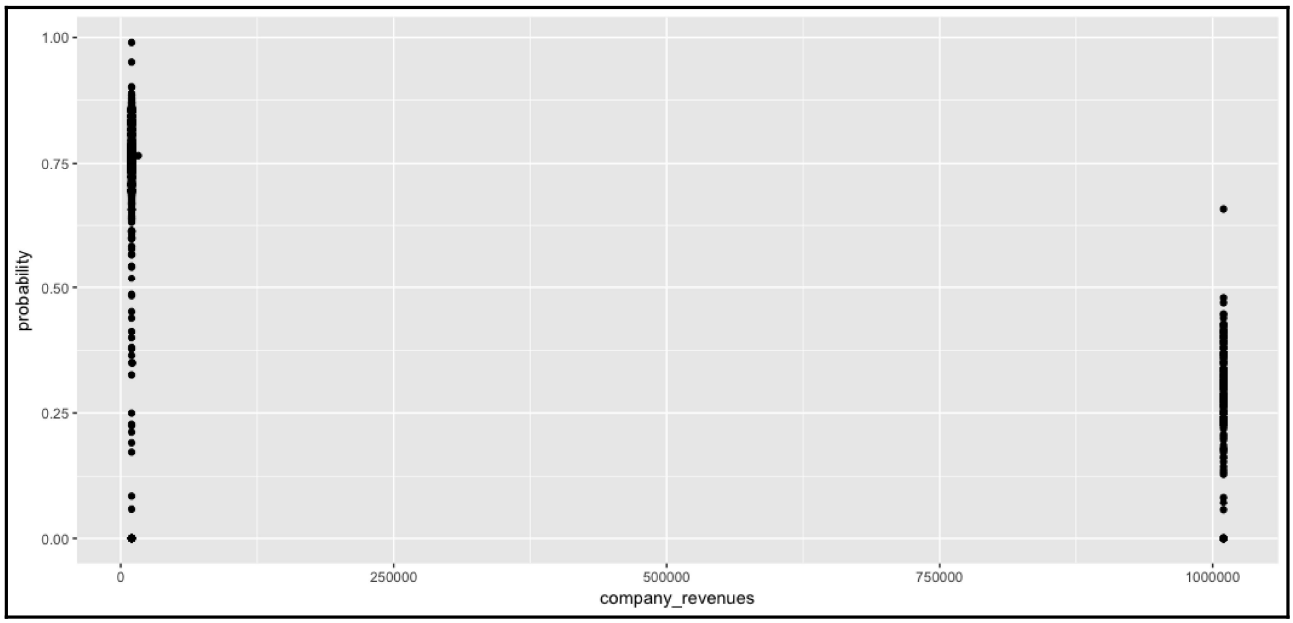

as.numeric(default_numeric)
R2
0.00
0.02
0.04
0.06
0.08
0
2
4
6
8
10
12
number of components

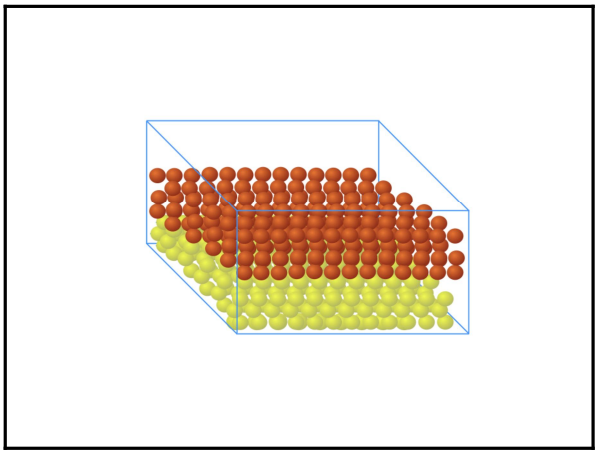
Chapter 10: A Different Outlook to Problems with Classification Models

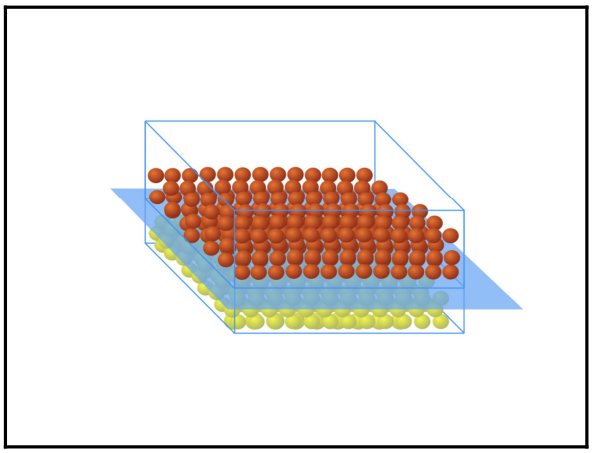


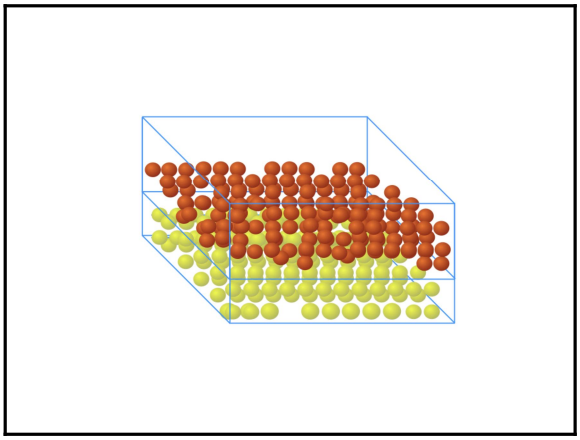

e^x
x

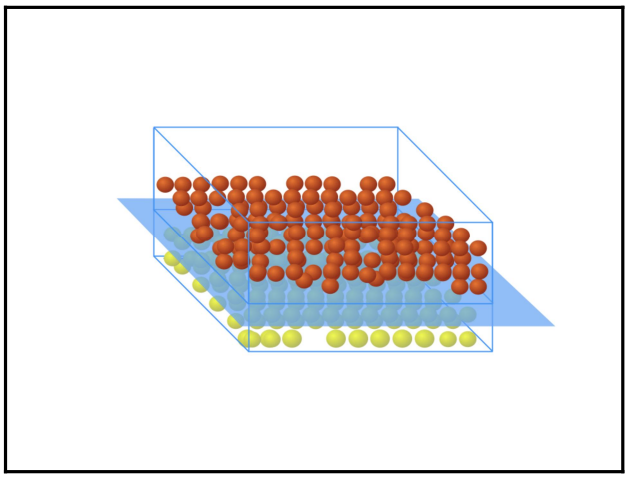


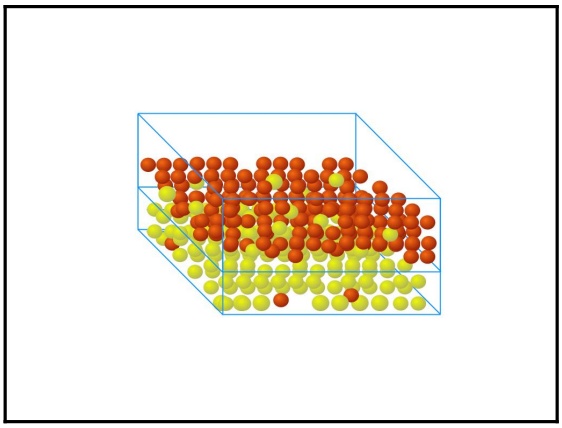

1.00
0.75
0.50
0.25
0.00
probability
0
250000
500000
750000
1000000
company_revenues

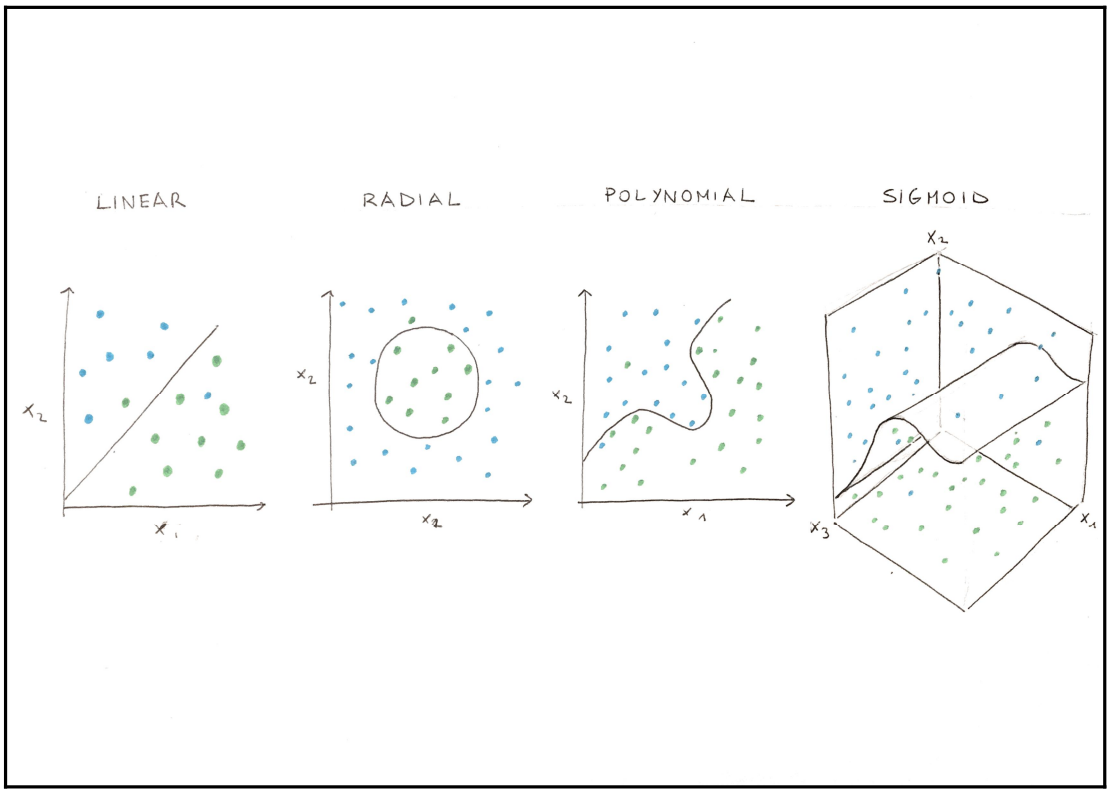




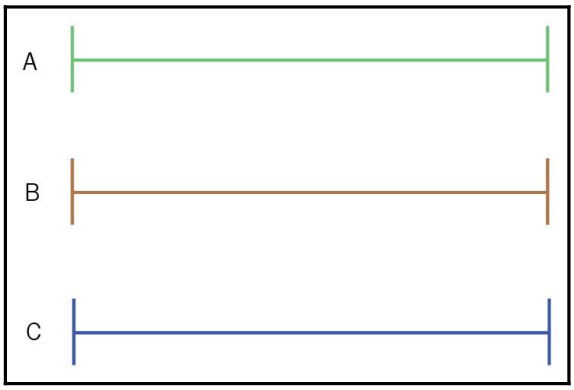


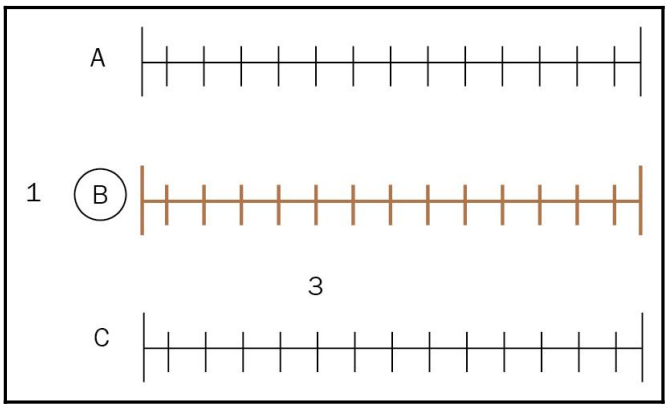


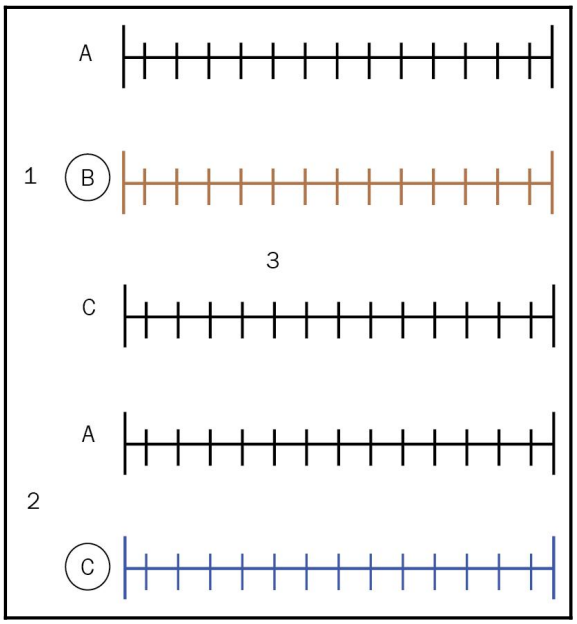


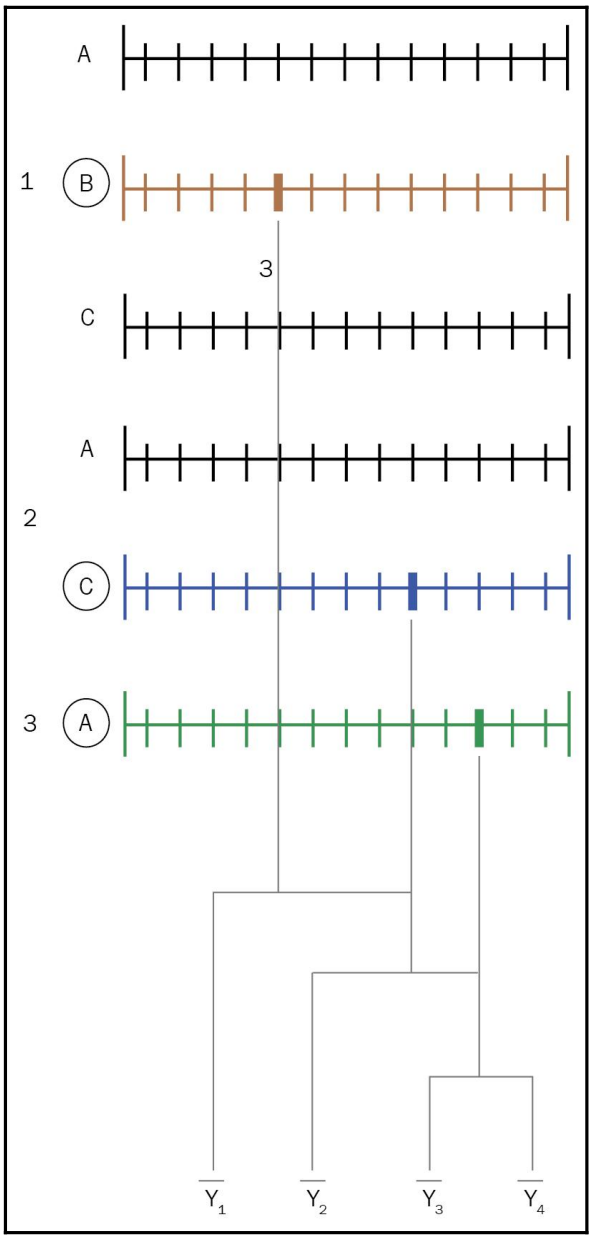

LINEAR
RADIAL
POLYNOMIAL
SIGMOID
X2
X1
X2
X1
X2
X1
X2
X3
X1

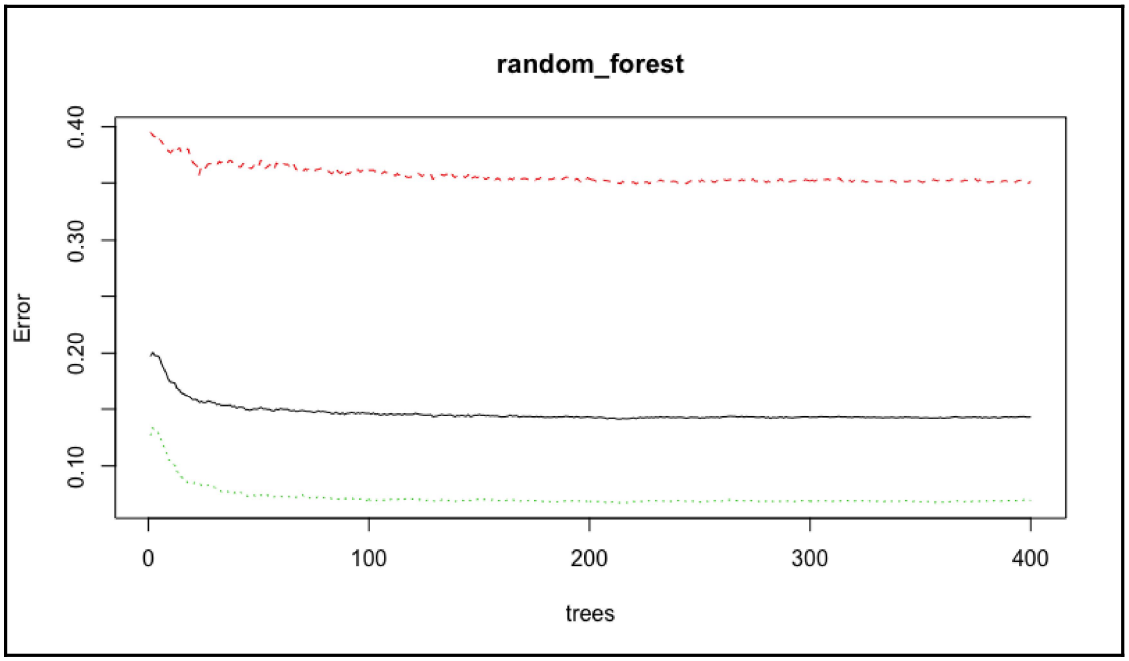
Chapter 11: The Final Clash – Random Forests and Ensemble Learning

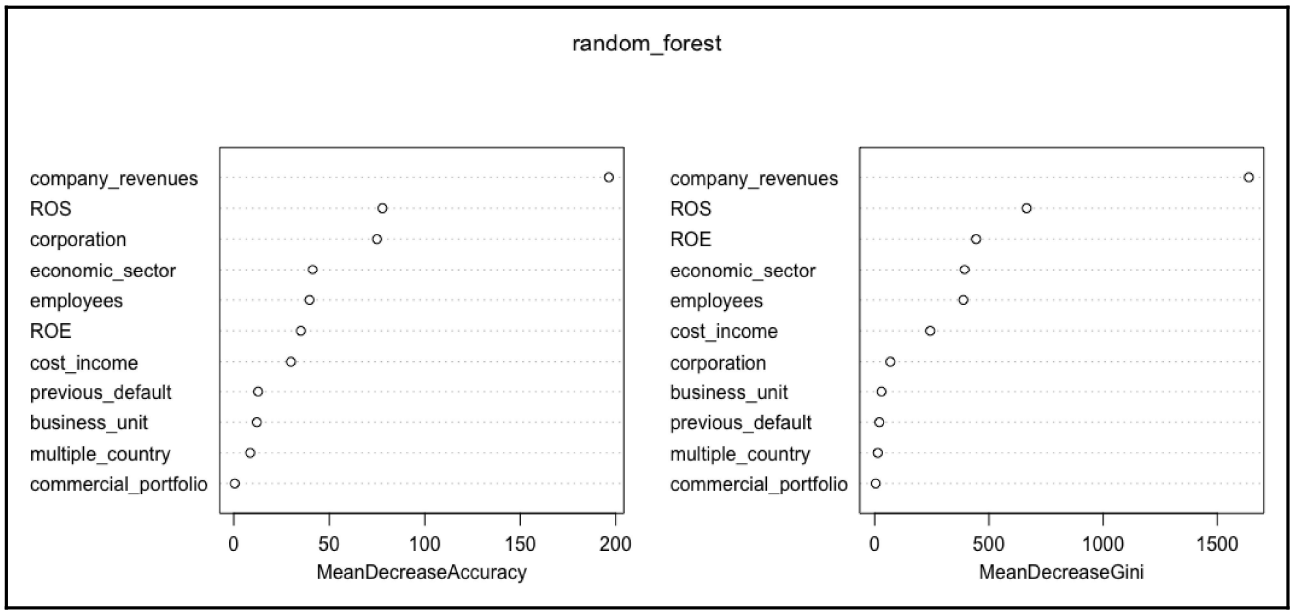


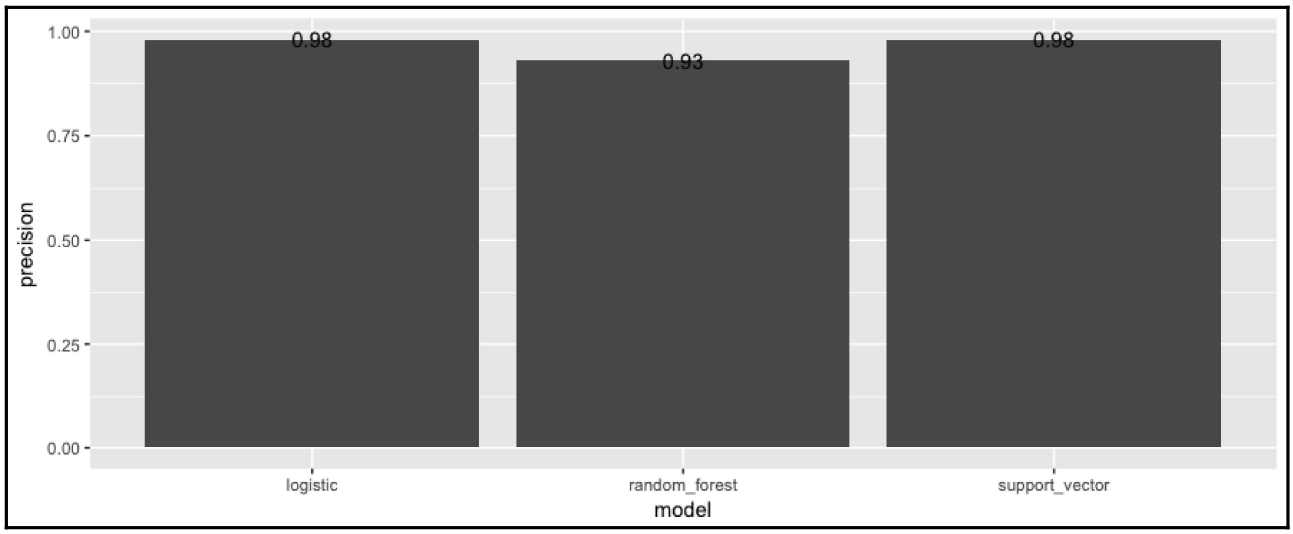



A
1
B
3
C
A
2
C

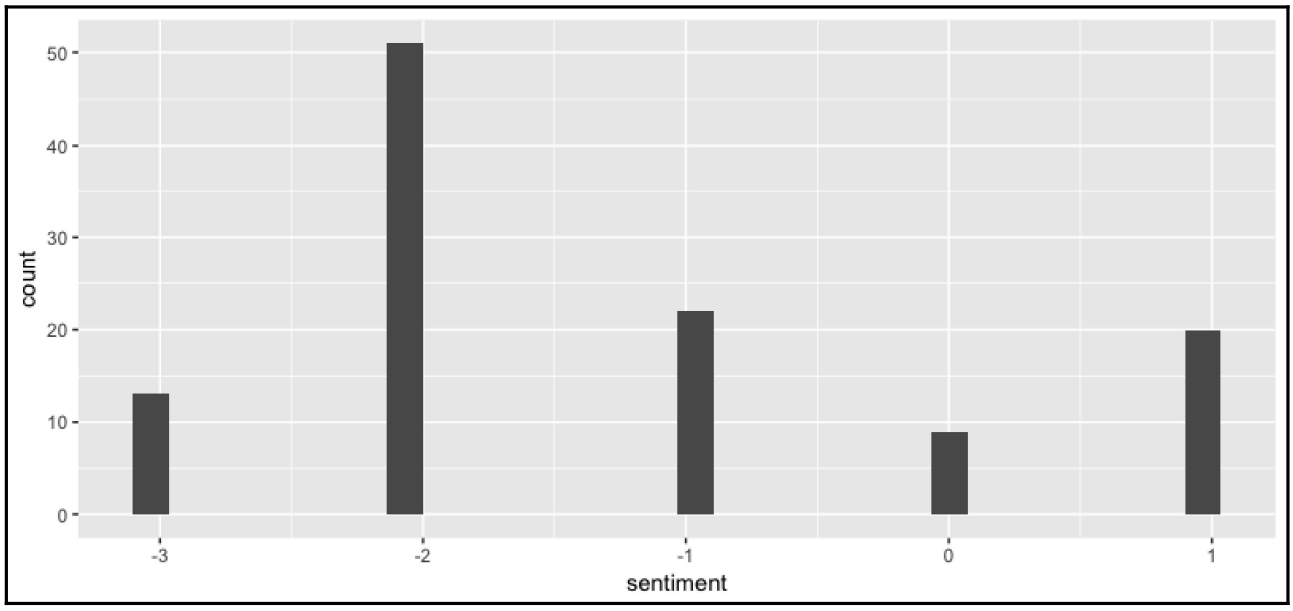

A
1
B
3
C
A
2
C
3
A
Y1
Y2
Y3
Y4

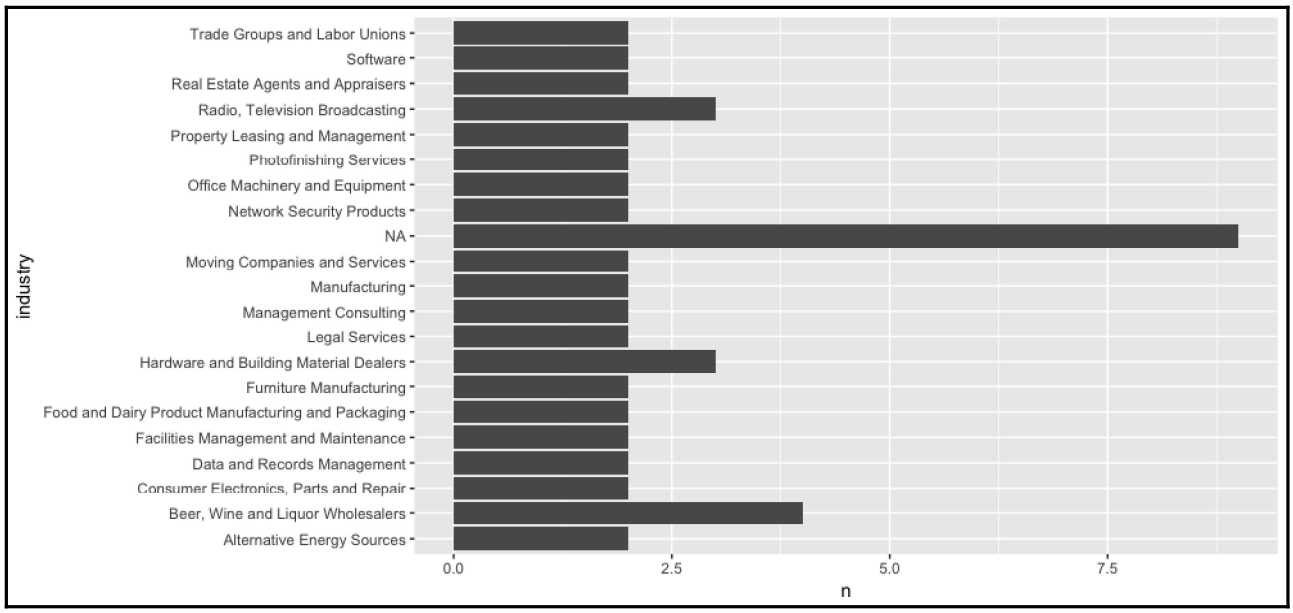

random_forest
Error
0.10
0.20
0.30
0.40
0
100
200
300
400
trees

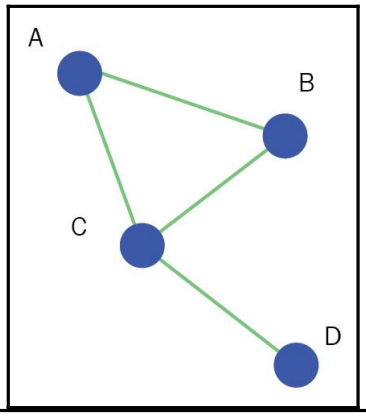

random_forest
company_revenues
ROS
corporation
economic_sector
employees
ROE
cost_income
previous_default
business_unit
multiple_country
commercial_portfolio
0
50
100
150
200
MeanDecreaseAccuracy
company_revenues
ROS
ROE
economic_sector
employees
cost_income
corporation
business_unit
previous_default
multiple_country
commercial_portfolio
0
500
1000
1500
MeanDecreaseGini

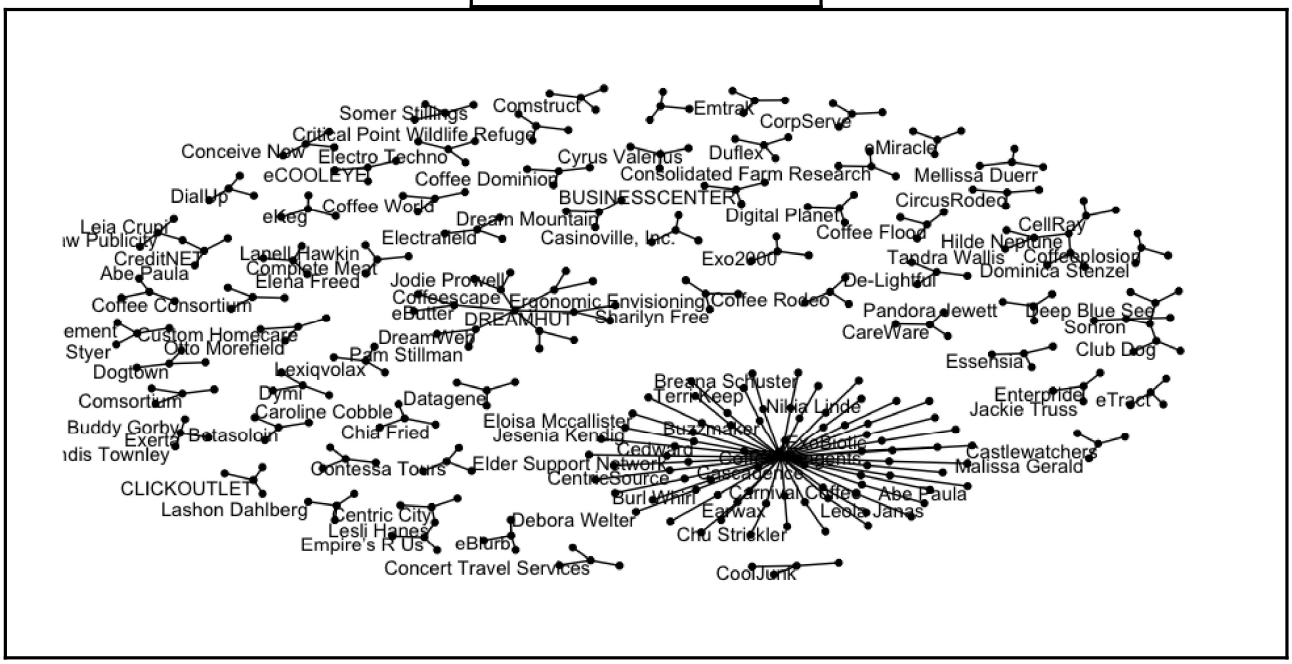

1.00
0.75
0.50
0.25
0.00
precision
0.98
0.93
0.98
logistic
random_forest
support_vector
model

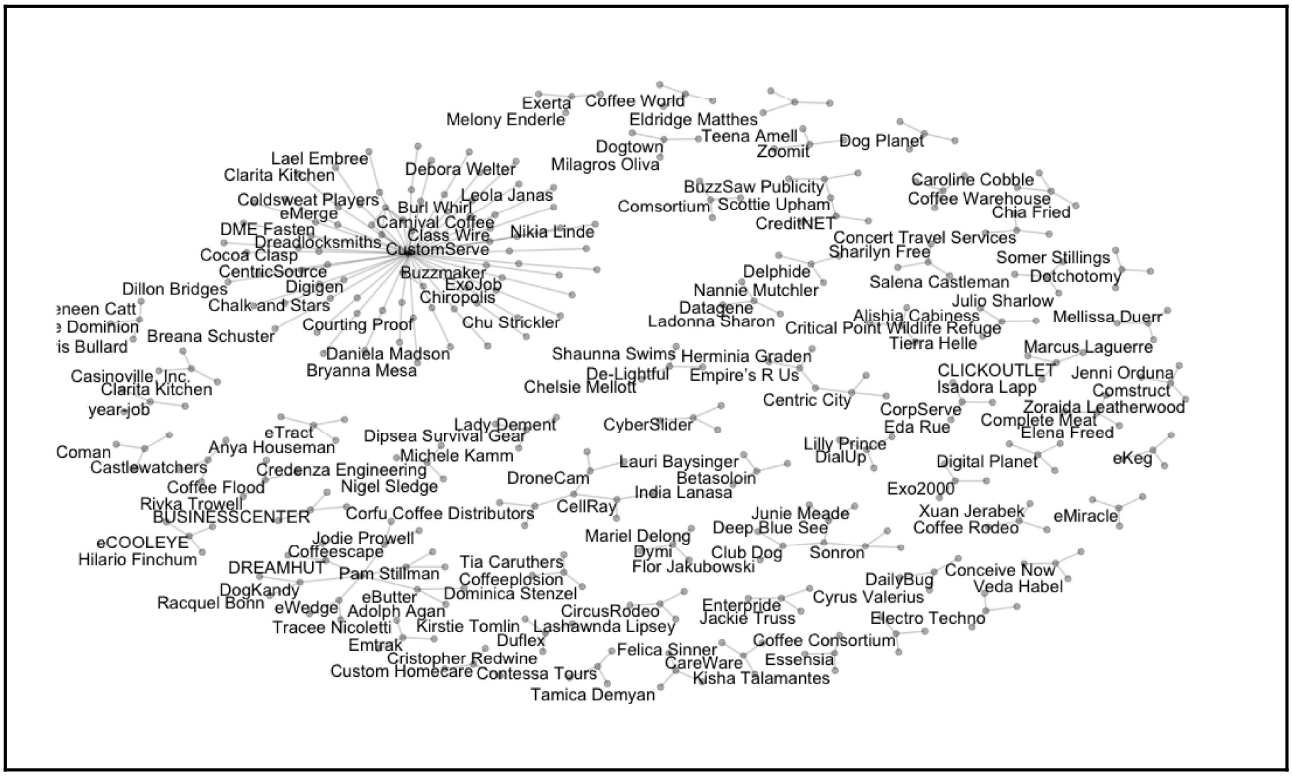
Chapter 12: Looking for the Culprit – Text Data Mining with R

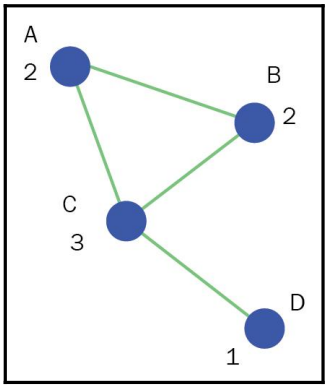


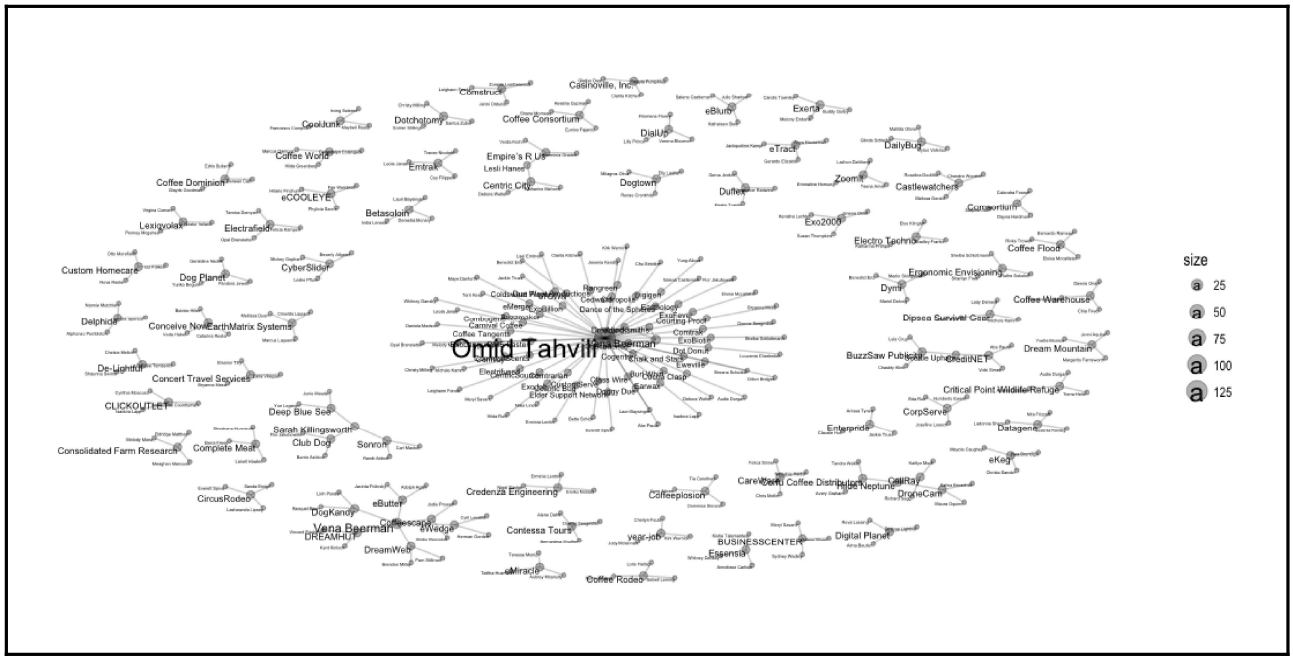



A
B
C
D


Somer Stillings
Comstruct
Emtrak
CorpServe
Critical Point Wildlife Refuge
Conceive Now
Electro Techno
Cyrus Valerius
Duflex
eMiracle
eCOOLEYE
Coffee Dominion
Consolidated Farm Research
Mellissa Duerr
DialUp
Coffee World
BUSINESSCENTER
CircusRodeo
eKeg
Digital Planet
Leia Crupi
Dream Mountain
Coffee Flood
CellRay
Electrafield
Casinoville, Inc.
Hilde Neptune
CreditNET
Lanell Hawkin
Exo2000
Tandra Wallis
Coffeeplosion
Abe Paula
Complete Meal
Elena Freed
Dominica Stenzel
Jodie Prowell
De-Lighted
Coffee Consortium
Coffeescape
Ergonomic Envisioning
Coffee Rodeo
Pandora Jewett
Deep Blue Sea
eButter
DREAMHUT
Sharilyn Free
Custom Homecare
CareWare
Sonron
Otto Morefield
Dream Web
Club Dog
Styer
Pam Stillman
Essensia
Dogtown
lexiqvolax
Breana Schuster
Terri Keep
Nikia Linde
Enterpride
eTract
Comsortium
Dymi
Datagene
Jackie Truss
Caroline Cobble
Eloisa Mccallister
Buddy Gorby
Chia Fried
Jesenia Kendig
Buzzmaker
Exerta
Betasolon
Cedwin
Castlewatchers
Elder Support Network
Malissa Gerald
Contessa Tours
CentreSource
CLICKOUTLET
Burl Whirl
Carnival Coffee
Abe Paula
Lashon Dahlberg
Centric City
Earwax
Leola Janas
Lesli Hanes
Debora Welter
Chu Strickler
Empire's R Us
eBlurb
Concert Travel Services
CoolJunk


Exerta
Coffee World
Melony Enderle
Eldridge Matthes
Teena Amell
Dog Planet
Zoomit
Dogtown
Milagros Oliva
Lael Embree
Clarita Kitchen
Debora Welter
Leola Janas
Burl Whirl
Coldsweat Players
eMerge
Carnival Coffee
Class Wire
Nikia Linde
DME Fasten
Dreadlocksmiths
CustomServe
Cocoa Clasp
CentricSource
Buzzmaker
Dillon Bridges
Digigen
ExoJob
Chiropolis
Chalk and Stars
Courting Proof
Chu Strickler
Breana Schuster
Daniela Madson
Bryanna Mesa
Casinoville Inc
Clarita Kitchen
year-job
BuzzSaw Publicity
Comsortium
Scottie Upham
CreditNET
Caroline Cobble
Coffee Warehouse
Chia Fried
Concert Travel Services
Sharilyn Free
Somer Stillings
Delphide
Nannie Mutchler
Datagene
Ladonna Sharon
Salena Castleman
Detchotomy
Julio Sharlow
Alishia Cabiness
Mellissa Duerr
Critical Point Wildlife Refuge
Tierra Helle
Marcus Laguerre
Shaunna Swims
Herminia Graden
De-Lightful
Empire's R Us
Chelsie Mellott
CLICKOUTLET
Jenni Orduna
Isadora Lapp
Comstruct
Centric City
CorpServe
Zoraida Leatherwood
Eda Rue
Complete Meat
Elena Freed
Lilly Prince
DialUp
Digital Planet
eKeg
Exo2000
Xuan Jerabek
eMiracle
Junie Meade
Deep Blue See
Coffee Rodeo
Mariel Delong
Dymi
Club Dog
Sonron
Flor Jakubowski
Conceive Now
DailyBug
Veda Habel
Cyrus Valerius
Enterpride
Jackie Truss
Electro Techno
Coffee Consortium
Felica Sinner
CareWare
Essensia
Kisha Talamantes
Tamica Demyan
Contessa Tours
Cristopher Redwine
Custom Homecare
Duflex
Emtrak
Tracee Nicoletti
Kirstie Tomlin
Lashawnda Lipsey
CircusRodeo
Adolph Agan
eWedge
eButter
Dominica Stenzel
Coffeeplosion
Tia Caruthers
Pam Stillman
DogKandy
Racquel Bonn
DREAMHUT
Coffeescape
Jodie Prowell
eCOOLEYE
Hilario Finchum
Rivka Trowell
BUSINESSCENTER
Corfu Coffee Distributors
CellRay
Coffee Flood
Nigel Sledge
DroneCam
India Lanasa
Betasoloin
Castlewatchers
Credenza Engineering
Michele Kamm
Lauri Baysinger
Coman
Anya Houseman
eTract
Dipsea Survival Gear
Lady Dement
CyberSlider
Ibneen Catt
Dominion
is Bullard


A
2
B
2
C
3
D
1


Omfd Tahvill
Vena Beerman
Castlewatchers
Zoomia
Exo2000
Electro Technics
Ergonomic Envisioning
Coffee Flood
Coffee Warehouse
Dipsea Survival Gear
Dream Mountain
BuzzSaw Publishing
CreditNET
Critical Point Wildlife Refuge
CorpServe
Enterprida
Datagen
eKeg
DroneCam
CellRay
Digital Planet
BUSINESSCENTER
Essensia
Coffee Rodeo
eMiracle
Contessa Tours
year-job
Credenza Engineering
Coffeeplosion
Carey Coffee Distributors
Blue Neptune
DreamWeb
DREAMHUT
DogKandy
CircusRodeo
Consolidated Farm Research
Complete Meal
Club Dog
Sorvon
Sarah Killingsworth
Deep Blue Sea
CLICKOUTLET
De-Lightful
Concert Travel Services
Conceive Now
EarthMatrix Systems
Delphida
Custom Homecare
Dog Planet
CyberSlider
Lexigrolab
Electrafield
Betasolain
eCOOLEYE
Coffee Dominion
Coffee World
CoolJunk
Dotchatomy
Construct
Cashoville, Inc.
Coffee Consortium
Empire's R Us
Emtrak
Centric City
Lesli Hanes
Dogtown
DialUp
eBluro
Exerra
eTrack
DailyBug
Duffel
Consortium
size
25
50
75
100
125