Index
Index

**Numbers & Symbols**
\ (backward slash) as separator, 69
/ (forward slash) as separator, 69
1-itemsets, 147
2-itemsets, 148–149
3 Vs (volume, variety, velocity), 2–3
3-itemsets, 149–150
4-itemsets, 150–151

**A**
accuracy, 225
ACF (autocorrelation function), 236–237
ACME text analysis example, 259–260
  raw text collection, 260–263
aggregates (SQL)
  ordered, 351–352
  user-defined, 347–351
aggregators of data, 18
AIE (Applied Information Economics), 28
algorithms
  clustering, 134–135
  decision trees, 197–200
    C4.5, 203–204
    CART, 204
    ID3, 203
Alphine Miner, 42
alternative hypothesis, 102–103
analytic projects
  Approach, 369–371
  BI analyst, 362
  business users, 361
  code, 362, 376–377
  communication, 360–361
  data engineer, 362
  data scientists, 362
  DBA (Database Administrator), 362
deliverables, 362–364
  audiences, 364–365
  core material, 364–365
key points, 372
Main Findings, 367–369
model description, 371
model details, 372–374
operationalizing, 360–361
outputs, 361
presentations, 362
Project Goals, 365–367
project manager, 362
project sponsor, 361
recommendations, 374–375
stakeholders, 361–362
  technical specifications, 376–377
analytic sandboxes. See sandboxes
analytical architecture, 13–15
analytics
  business drivers, 11
  examples, 22–23
    new approaches, 16–19
ANOVA, 110–114
Anscombe’s quartet, 82–83
aov( ) function, 78
Apache Hadoop. See Hadoop
APIs (application programming interfaces), Hadoop, 304–305
apriori( ) function, 146, 152–157
Apriori algorithm, 139
grocery store example, 143
  Groceries dataset, 144–146
  itemset generation, 146–151
  rule generation, 152–157
items, 139, 140–141
  counting, 158
  partitioning and, 158
  sampling and, 158
  transaction reduction and, 158
architecture, analytical, 13–15
arima( ) function, 246
ARIMA (Autoregressive Integrated Moving Average) model, 236
  ACF, 236–237
  ARMA model, 241–244
  autoregressive models, 238–239
    building, 244–252
    cautions, 252–253
    constant variance, 250–251
    evaluating, 244–252
    fitted time series models, 249–250
    forecasting, 251–252
    moving average models, 239–241
    normality, 250–251
    PACF, 238–239
    reasons to choose, 252–253
    seasonal autoregressive integrated moving average
      model, 243–244
    VARIMA, 253
ARMA (Autoregressive Moving Average) model, 241–244
array( ) function, 74
arrays
  matrices, 74
  R, 74–75
association rules, 138–139
  application, 143
  candidate rules, 141–142
diagnostics, 158
testing and, 157–158
validation, 157–158
attributes
objects, k-means, 130–131
R, 71–72
AUC (area under the curve), 227
autoregressive models, 238–239
averages, moving average models, 239–241

B
bagging, 228
bag-of-words in text analysis, 265–266
banking, 18
barplot( ) function, 88
barplots, 93–94
Bayes' Theorem, 212–214. See also naïve Bayes
conditional probability, 212
BI (business intelligence)
analytical tools, 10
versus Data Science, 12–13
Big Data
3 Vs, 2–3
analytics, examples, 22–23
characteristics, 2
definitions, 2–3
drivers, 15–16
ecosystem, 16–19
key roles, 19–22
McKinsey & Co. on, 3
volume, 2–3
boosting, 228–229
bootstrap aggregation, 228
box-and-whisker plots, 95–96
Box-Jenkins methodology, 235–236
ARIMA model, 236
branches (decision trees), 193
Brown Corpus, 267–268
business drivers for analytics, 11
Business Intelligence Analyst, Operationalize phase, 52
Business Intelligence Analyst role, 27
Business User, Operationalize phase, 52
Business User role, 27
buyers of data, 18

C
C4.5 algorithm, 203–204
cable TV providers, 17
candidate rules, 141–142
CART (Classification And Regression Trees), 204
case folding in text analysis, 264–265
categorical algorithms, 205
categorical variables, 170–171
cbind( ) function, 78
centroids, 120–122
starting positions, 134
character data types, R, 72
charts, 386–387
churn rate (customers), 120
logistic regression, 180–181
class( ) function, 72
classification
bagging, 228
boosting, 228–229
bootstrap aggregation, 228
decision trees, 192–193
algorithms, 197–200, 203–204
binary decisions, 206
branches, 193
categorical attributes, 205
classification trees, 193
correlated variables, 206
decision stump, 194
evaluating, 204–206
greedy algorithm, 204
internal nodes, 193
irrelevant variables, 205
nodes, 193
numerical attributes, 205
R and, 206–211
redundant variables, 206
regions, 205
regression trees, 193
root, 193
short trees, 194
splits, 193, 194, 197, 200–203
structure, 205
uses, 194
naïve Bayes, 211–212
Bayes' theorem, 212–214
diagnostics, 217–218
naïve Bayes classifier, 214–217
R and, 218–224
smoothing, 217
classification trees, 193
classifiers
accuracy, 225
diagnostics, 224–228
recall, 225
clickstream, 9
clustering, 118
algorithms, 134–135
centroids, 120–122
starting positions, 134

diagnostics, 128–129

k-means, 118–119

algorithm, 120–122
customer segmentation, 120
image processing and, 119
medical uses, 119
reasons to choose, 130–134
rescaling, 133–134
units of measure, 132–133

labels, 127

number of clusters, 123–127
code, technical specifications in project, 376–377
coefficients, linear regression, 169
combiners, 302–303
Communicate Results phase of lifecycle, 30, 49–50
components, short trees as, 194
conditional entropy, 199
conditional probability, 212
naïve Bayes classifier, 215–216
confidence, 141–142
outcome, 172
parameters, 171
confidence interval, 107
confint( ) function, 171
confusion matrix, 224, 280
contingency tables, 79
continuous variables, discretization, 211
corpora
Brown Corpus, 267–268
corpora in Natural Language Processing, 256
IC (information content), 268–269
sentiment analysis and, 278
correlated variables, 206
credit card companies, 2
CRISP-DM, 28
crowdsourcing, 17
CSV (comma-separated-value) files, 64–65
importing, 64–65
customer segmentation
k-means, 120
logistic regression, 180–181
CVS files, 6
cyclic components of time series analysis, 235

data

growth needs, 9–10
sources, 15–16
data( ) function, 84
data aggregators, 17–18
data analysis, exploratory, 80–82
visualization and, 82–85

Data Analytics Lifecycle

Business Intelligence Analyst role, 27
Business User role, 27
Communicate Results phase, 30, 49–50
GINA case study, 58–59
Data Engineer role, 27–28
Data preparation phase, 29, 36–37
Alpine Miner, 42
data conditioning, 40–41
data visualization, 41–42
Data Wrangler, 42
dataset inventory, 39–40
ETLT, 38–39
GINA case study, 55–56
Hadoop, 42
OpenRefine, 42
sandbox preparation, 37–38
tools, 42

Data Scientist role, 28

DBA (Database Administrator) role, 27

Discovery phase, 29
business domain, 30–31
data source identification, 35–36
framing, 32–33
GINA case study, 54–55
hypothesis development, 35
resources, 31–32
sponsor interview, 33–34
stakeholder identification, 33

GINA case study, 53–60

Model Building phase, 30, 46–48
Alpine Miner, 48
GINA case study, 56–58
Mathematica, 48
Matlab, 48
Octave, 48
PL/R, 48
Python, 48
R, 48
SAS Enterprise Miner, 48
SPSS Modeler, 48
SQL, 48
STATISTICA, 48
WEKA, 48

Model Planning phase, 29–30, 42–44
data exploration, 44–45
GINA case study, 56
model selection, 45
R, 45–46
SAS/ACCESS, 46
SQL Analysis services, 46
variable selection, 44–45
Operationalize phase, 30, 50–53, 360
  Business Intelligence Analyst and, 52
  Business User and, 52
  Data Engineer and, 52
  Data Scientist and, 52
  DBA (Database Administrator) and, 52
  GINA case study, 59–60
  Project Manager and, 52
  Project Sponsor and, 52
processes, 28
  Project Manager role, 27
  Project Sponsor role, 27
roles, 26–28
data buyers, 18
data cleansing, 86
data collectors, 17
data conditioning, 40–41
data creation rate, 3
data devices, 17
Data Engineer, Operationalize phase, 52
Data Engineer role, 27–28
data formats, text analysis, 257
data frames, 75–76
data marts, 10
Data preparation phase of lifecycle, 29, 36–37
data conditioning, 40–41
data visualization, 41–42
  dataset inventory, 39–40
  ETLT, 38–39
  sandbox preparation, 37–38
data repositories, 9–11
types, 10–11
Data Savvy Professionals, 20
Data Science versus BI, 12–13
Data Scientists, 28
  activities, 20–21
  business challenges, 20
  characteristics, 21–22
  Operationalize phase and, 52
  recommendations and, 21
  statistical models and, 20–21
data sources
  Discovery phase, 35–36
  text analysis, 257
data structures, 5–9
  quasi-structured data, 6, 7
  semi-structured data, 6
  structured data, 6
  unstructured data, 6
data types in R, 71–72
  character, 72
  logical, 72
  numeric, 72
  vectors, 73–74
data users, 18
data visualization, 41–42, 377–378
  CSS and, 378
  Ggobi, 377–378
  Gnuplot, 377–378
  graphs, 380–386
    clean up, 387–392
    three-dimensional, 392–393
  HTML and, 378
  key points with support, 378–379
  representation methods, 386–387
  SVG and, 378
data warehouses, 11
Data Wrangler, 42
datasets
  exporting, R and, 69–71
  importing, R and, 69–71
  inventory, 39–40
Davenport, Tom, 28
DBA (Database Administrator), 10, 27
  Operational phase and, 52
decision trees, 192–193
  algorithms, 197–200
    C4.5, 203–204
    CART, 204
    categorical, 205
    greedy, 204
    ID3, 203
    numerical, 205
  binary decisions, 206
  branches, 193
  classification trees, 193
  correlated variables, 206
  evaluating, 204–206
  greedy algorithms, 204
  internal nodes, 193
  irrelevant variables, 205
  nodes
    depth, 193
    leaf, 193
  R and, 206–211
  redundant variables, 206
  regions, 205
  regression trees, 193
  root, 193
  short trees, 194
  decision stump, 194
Index

splits, 193, 197
detecting, 200–203
limiting, 194
structure, 205
uses, 194
Deep Analytical Talent, 19–20
DELTA framework, 28
demand forecasting, linear regression and, 162
density plots, exploratory data analysis, 88–91
dependent variables, 162
descriptive statistics, 79–80
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
deviance, 183–184
development of logistic regression, 174
linear regression
linearity assumption, 173
N-fold cross-validation, 177–178
normality assumption, 174–177
residuals, 173–174
logistic regression
deviance, 183–184
histogram of probabilities, 188
log-likelihood test, 184–185
pseudo-R², 183
ROC curve, 185–187
naïve Bayes, 217–218
diff( ) function, 245
difference in means, 104
confidence interval, 107
student’s t-testing, 104–106
Welch’s t-test, 106–108
differencing, 241–242
dirty data, 85–87
Discovery phase of lifecycle, 29
data source identification, 35–36
framing, 32–33
hypothesis development, 35
sponsor interview, 33–34
stakeholder identification, 33
Discretization of continuous variables, 211
documents, categorization, 274–277
dotchart( ) function, 88

E

Eclipse, 304
ecosystem of Big Data, 16–19

Data Savvy Professionals, 20
Deep Analytical Talent, 19–20
key roles, 19–22
Technology and Data Enablers, 20
EDWs (Enterprise Data Warehouses), 10
effect size, 110
EMC Google search example, 7–9
emoticons, 282
engineering, logistic regression and, 179
ensemble methods, decision trees, 194
error distribution
linear regression model, 165–166
residual standard error, 170
ETLT, 38–39
EXCEPT operator (SQL), 333–3334
exploratory data analysis, 80–82
density plot, 88–91
dirty data, 85–87
histograms, 88–91
multiple variables, 91–92
analysis over time, 99
barplots, 93–94
box-and-whisker plots, 95–96
dotcharts, 93–94
hexbinplots, 96–97
versus presentation, 99–101
scatterplot matrix, 97–99
visualization and, 82–85
single variable, 88–91
exporting datasets in R, 69–71
expressions, regular, 263

F

Facebook, 2, 3–4
factors, 77–78
financial information, logistic regression and, 179
FNR (false negative rate), 225
forecasting
ARIMA (Autoregressive Integrated Moving Average)
model, 251–252
linear regression and, 162
FP (false positives), confusion matrix, 224
FPR (false positive rate), 225
framing in Discovery phase, 32–33
functions
aov( ), 78
apriori( ), 146, 152–157
arima( ), 246
array( ), 74
barplot( ), 88
cbind( ), 78
class( ), 72
confint( ), 171
data( ), 84
diff( ), 245
dotchart( ), 88
gl( ), 84
glm( ), 183
hclust( ), 135
head( ), 65
inspect( ), 147, 154–155
integer( ), 72
IQR( ), 80
is.data.frame( ), 75
is.na( ), 86
is.vector( ), 73
jpeg( ), 71
kmeans( ), 134
kmode( ), 134–135
length( ), 72
library( ), 70
lm( ), 66
load.image( ), 68–69
matrix.inverse( ), 74
mean( ), 86
my_range( ), 80
na.exclude( ), 86
pamk( ), 135
Pig, 307–308
plot( ), 65, 65–154, 154, 245
predict( ), 172
rbind( ), 78
read.csv( ), 64–65, 75
read.csv2( ), 70
read.delim2( ), 70
rpart, 207
SQL, 347–351
sqlQuery( ), 70
str( ), 75
summary( ), 65, 66–67, 79, 80–82
t( ), 74
ts( ), 245
typeof( ), 72
wilcox.test( ), 109
window functions (SQL), 343–347
write.csv( ), 70
write.csv2( ), 70
write.table( ), 70

G

Generalized Linear Model function, 182
genetic sequencing, 3, 4
genomics, 4, 16
genotyping, 4
GGobi, 377–378

GINA (Global Innovation Network and Analysis), Data Analytics Lifecycle case study, 53–60
gl( ) function, 84
glm( ) function, 183
Gnuplot, 377–378
GPS systems, 16
Graph Search (Facebook), 3–4
graphs, 380–386
clean up, 387–392
three-dimensional, 392–393
greedy algorithms, 204
Green Eggs and Ham, text analysis and, 256
grocery store example of Apriori algorithm, 143
Groceries dataset, 144–146
itemsets, frequent generation, 146–151
rules, generating, 152–157
growth needs of data, 9–10
GUIs (graphical user interfaces), R and, 67–69

H

Hadoop
Data preparation phase, 42
Hadoop Streaming API, 304–305
HBase, 311–312
architecture, 312–317
column family names, 319
column qualifier names, 319
data model, 312–317
Java API and, 319
rows, 319
use cases, 317–319
versioning, 319
Zookeeper, 319
HDFS, 300–301
Hive, 308–311
LinkedIn, 297
Mahout, 319–320
MapReduce, 22
combiners, 302–303
development, 304–305
drivers, 301
execution, 304–305
mappers, 301–302
partitioners, 304
structuring, 301–304
natural language processing, 18
Pig, 306–308
pipes, 305
Watson (IBM), 297
Yahoo!, 297–298
YARN (Yet Another Resource Negotiator), 305
hash-based itemsets, Apriori algorithm and, 158
HAWQ (Hadoop With Query), 321
HBase, 311–312
  architecture, 312–317
  column family names, 319
  column qualifier names, 319
  data model, 312–317
  Java API and, 319
  rows, 319
  use cases, 317–319
  versioning, 319
  Zookeeper, 319
hclust( ) function, 135
HDFS (Hadoop Distributed File System), 300–301
head( ) function, 65
hexbinplots, 96–97
histograms
  exploratory data analysis, 88–91
  logistic regression, 188
Hive, 308–311
HiveQL (Hive Query Language), 308
Hopper, Grace, 299
Hubbard, Doug, 28
HVE (Hadoop Virtualization Extensions), 321
hypotheses
  alternative hypothesis, 102–103
  Discovery phase, 35
  null hypothesis, 102
hypothesis testing, 102–104
  two-sided hypothesis testing, 105
  type I errors, 109–110
  type II errors, 109–110
IBM Watson, 297
ID3 algorithm, 203
IDE (Interactive Development Environment), 304
IDF (inverted document frequency), 271–272
importing datasets in R, 69–71
in-database analytics
  SQL, 328–338
  text analysis, 338–339
independent variables, 162
input variables, 192
inspect( ) function, 147, 154–155
integer( ) function, 72
internal nodes (decision trees), 193
Internet of Things, 17–18
INTERSECT operator (SQL), 333
IQR( ) function, 80
is.data.frame( ) function, 75
is.na( ) function, 86
is.vector( ) function, 73
itemsets, 139
  1-itemsets, 147
  2-itemsets, 148–149
  3-itemsets, 149–150
  4-itemsets, 150–151
  Apriori algorithm, 139
  Apriori property, 139
downward closure property, 139
dynamic counting, Apriori algorithm and, 158
frequent itemset, 139
generation, frequent, 146–151
hash-based, Apriori algorithm and, 158
k-itemset, 139, 140–141
joins (SQL), 330–332
jpeg( ) function, 71
k clusters
  finding, 120–122
  number of, 123–127
  k-itemset, 139, 140–141
  k-means, 118–119
  customer segmentation, 120
  image processing and, 119
  k clusters
    finding, 120–122
    number of, 123–127
  medical uses, 119
  objects, attributes, 130–131
  R and, 123–127
  reasons to choose, 130–134
  rescaling, 133–134
  units of measure, 132–133
kmeans( ) function, 134
kmode( ) function, 134–135
lag, 237
Laplace smoothing, 217
lasso regression, 189
LDA (latent Dirichlet allocation), 274–275
leaf nodes, 192, 193
lemmatization, text analysis and, 258
length( ) function, 72
leverage, 142
library( ) function, 70
lifecycle. See also Data Analytics Lifecycle
lift, 142
linear regression, 162
coefficients, 169
diagnostics
  linearity assumption, 173
  N-fold cross-validation, 177–178
  normality assumption, 174–177
  residuals, 173–174
model, 163–165
categorical variables, 170–171
normally distributed errors, 165–166
outcome confidence intervals, 172
parameter confidence intervals, 171
prediction interval on outcome, 172
R, 166–170
p-values, 169–170
use cases, 162–163
LinkedIn, 2, 22–23, 297
lists in R, 76–77
lm( ) function, 66
load.image( ) function, 68–69
logical data types, R, 72
logistic regression, 178
cautions, 188–189
diagnostics, 181–182
deviance, 183–184
histogram of probabilities, 188
log-likelihood test, 184–185
pseudo-R², 183
ROC curve, 185–187
Generalized Linear Model function, 182
model, 179–181
multinomial, 190
reasons to choose, 188–189
use cases, 179
log-likelihood test, 184–185
loyalty cards, 17

M
MAD (Magnetic/Agile/Deep) skills, 28, 352–356
MADlib, 352–356
Mahout, 319–320
MapReduce, 22, 298–299
  combinators, 302–303
  development, 304–305
  drivers, 301–302
  execution, 304–305
  mappers, 301–302
  partitioners, 304
  structuring, 301–304
market basket analysis, 139
  association rules, 143
marketing, logistic regression and, 179
master nodes, 301
matrices
  confusion matrix, 224
  R, 74–75
  scatterplot matrices, 97–99
matrix.inverse( ) function, 74
MaxEnt (maximum entropy), 278
McKinsey & Co. definition of Big Data, 3
mean( ) function, 86
medical information, 16
  k-means and, 119
  linear regression and, 162
  logistic regression and, 179
minimum confidence, 141
missing data, 86
mobile devices, 16
mobile phone companies, 2
Model Building phase of lifecycle, 30, 46–48
  Alpine Miner, 48
  Mathematica, 48
  Matlab, 48
  Octave, 48
  PL/R, 48
  Python, 48
  R, 48
  SAS Enterprise Miner, 48
  SPSS Modeler, 48
  SQL, 48
  STATISTICA, 48
  WEKA, 48
Model Planning phase of lifecycle, 29–30, 42–44
  data exploration, 44–45
  model selection, 45
  R, 45–46
  SAS/ACCESS, 46
  SQL Analysis services, 46
  variables, selecting, 44–45
morphological features in text analysis, 266–267
moving average models, 239–241
MPP (massively parallel processing), 5
MTurk (Mechanical Turk), 282
multinomial logistic regression, 190
multivariate time series analysis, 253
my_range( ) function, 80

N
na.exclude( ) function, 86
naive Bayes, 211–212
  Bayes’ theorem, 212–214
diagnostics, 217–218
naïve Bayes classifier, 214–217
R and, 218–224
sentiment analysis and, 278
smoothing, 217
natural language processing, 18
N-fold cross-validation, 177–178
NLP (Natural Language Processing), 256
nodes
master, 301
worker, 301
nodes (decision trees), 192
depth, 193
leaf, 193
leaf nodes, 192, 193
nonparametric tests, 108–109
nontraditional devices, 16
normality
ARIMA model, 250–251
linear regression, 174–177
normalization, data conditioning, 40–41
NoSQL, 322–323
null deviance, 183
null hypothesis, 102
numeric data types, R, 72
numerical algorithms, 205
numerical underflow, 216–217

O
objects, k-means, attributes, 130–131
OLAP (online analytical processing), 6
cubes, 10
OpenRefine, 42
Operationalize phase of lifecycle, 30, 50–53, 360
Business Intelligence Analyst and, 52
Business User and, 52
Data Engineer and, 52
Data Scientist and, 52
DBA (Database Administrator) and, 52
Project Manager and, 52
Project Sponsor and, 52
operators, subsetting, 75
outcome
confidence intervals, 172
prediction interval, 172

P
PACF (partial autocorrelation function), 238–239
pamk( ) function, 135
parameters, confidence intervals, 171
parametric tests, 108–109

R
arrays, 74–75
attributes, types, 71–72
data frames, 75–76
data types, 71–72
character, 72
logical, 72
numeric, 72
vectors, 73–74
decision trees, 206–211
descriptive statistics, 79–80
exploratory data analysis, 80–82
density plot, 88–91
dirty data, 85–87
histograms, 88–91
multiple variables, 91–99
versus presentation, 99–101
visualization and, 82–85, 88–91
factors, 77–78
functions
aov( ), 78
array( ), 74
barplot( ), 88
cbind( ), 78
class( ), 72
data( ), 84
dotchart( ), 88
gl( ), 84
head( ), 65
import function defaults, 70
integer( ), 72
IQR( ), 80
is.data.frame( ), 75
is.na( ), 86
is.vector( ), 73
jpeg( ), 71
length( ), 72
library( ), 70
lm( ), 66
load.image( ), 68–69
my_range( ), 80
plot( ) function, 65
rbind( ), 78
read.csv( ), 65, 75
read.csv2( ), 70
read.delim( ), 69
read.delim2( ), 70
read.table( ), 69
str( ), 75
summary( ), 65, 66–67, 79
t( ), 74
typeof( ), 72
visualizing single variable, 88
write.csv( ), 70
write.csv2( ), 70
write.table( ), 70

GUIs, 67–69
import/export, 69–71
k-means analysis, 123–127
linear regression model, 166–170
lists, 76–77
matrices, 74–75
model planning and, 45–46
naïve Bayes and, 218–224
operators, subsetting, 75
overview, 64–67
statistical techniques, 101–102
ANOVA, 110–114
difference in means, 104–108
effect size, 110
hypothesis testing, 102–104
power of test, 110
sample size, 110
type I errors, 109–110
type II errors, 109–110
tables, contingency tables, 79
R commander GUI, 67
random components of time series analysis, 235
Rattle GUI, 67
raw text
collection, 260–263
tokenization, 264
rbind( ) function, 78
RDBMS, 6
read.csv( ) function, 64–65, 75
read.csv2( ) function, 70
read.delim( ) function, 69
read.delim2( ) function, 70
read.table( ) function, 69
real estate, linear regression and, 162
recall in sentiment analysis, 281
redundant variables, 206
regression
lasso, 189
linear, 162
coefficients, 169
diagnostics, 173–178
model, 163–172
p-values, 169–170
use cases, 162–163
logistic, 178
cautions, 188–189
diagnostics, 181–188
model, 179–181
multinomial logistic, 190
reasons to choose, 188–189
use cases, 179
multinomial logistic, 190
ridge, 189
variables
dependent, 162
independent, 162
regression trees, 193
regular expressions, 263, 339–340
relationships, 141
repositories, 9–11
types, 10–11
representation methods, 386–387
rescaling, k-means, 133–134
residual deviance, 183
residual standard error, 170
residuals, linear regression, 173–174
resources, Discovery phase of lifecycle, 31–32
RFID readers, 16
ridge regression, 189
ROC (receiver operating characteristic) curve, 185–187, 225
roots (decision trees), 193
rpart function, 207
RStudio GUI, 67–68
rules
  association rules, 138–139
  application, 143
  candidate rules, 141–142
  diagnostics, 158
  testing and, 157–158
  validation, 157–158
  generating, grocery store example (Apriori), 152–157

S
sales, time series analysis and, 234
sample size, 110
sampling, Apriori algorithm and, 158
sandboxes, 10, 11. See also work spaces
  Data preparation phase, 37–38
SAS/ACCESS, model planning, 46
scatterplot matrix, 97–99
scatterplots, 81
  Anscombe's quartet, 83
  multiple variables, 91–92
scientific method, 28
searches, text analysis and, 257
seasonal autoregressive integrated moving average model, 243–244
seasonality components of time series analysis, 235
seismic processing, 16
semi-structured data, 6
SensorNet, 17–18
sentiment analysis in text analysis, 277–283
  confusion matrix, 280
  precision, 281
  recall, 281
shopping
  loyalty cards, 17
  RFID chips in carts, 17
short trees, 194
smart devices, 16
smartphones, 17
smoothing, 217
social media, 3–4
sources of data, 15–16
spart parts planning, time series analysis and, 234–235
splits (decision trees), 193
detecting, 200–203
sponsor interview, Discovery phase, 33
spreadmarts, 10
spreadsheets, 6, 9, 10
SQL (Structured Query Language), 328–329
  aggregates
    ordered, 351–352
    user-defined, 347–351
  EXCEPT operator, 333–3334
  functions, user-defined, 347–351
  grouping, 334–338
  INTERSECT operator, 333
  joins, 330–332
  MADlib, 352–356
  queries, 329–330
    nested, 3334
    subqueries, 3334
  set operations, 332–334
  UNION ALL operator, 332–333
  window functions, 343–347
SQL Analysis services, model planning and, 46
sqlQuery( ) function, 70
stakeholders, Discovery phase of lifecycle, 33
stationary time series, 236
statistical techniques, 101–102
  ANOVA, 110–114
  difference in means, 104
    student's t-test, 104–106
    Welch's t-test, 106–108
  effect size, 110
  hypothesis testing, 102–104
  power of test, 110
  sample size, 110
  type I errors, 109–110
  type II errors, 109–110
  Wilcoxon rank-sum test, 108–109
statistics
  Anscombe’s quartet, 82–83
  descriptive, 79–80
stemming, text analysis and, 258
stock trading, time series analysis and, 235
stop words, 270–271
str( ) function, 75
structured data, 6
subsetting operators, 75
summary( ) function, 65, 66–67, 79, 80–82
SVM (support vector machines), 278

T
t( ) function, 74
tables, contingency tables, 79
Target stores, 22
t-distribution
ANOVA, 110–114
students’ t-test, 104–106
Welch’s t-test, 106–108
technical specifications in project, 376–377
Technology and Data Enablers, 20
testing, association rules and, 157–158
text analysis, 256
ACME example, 259–263
bag-of-words, 265–266
corpora, 264–265
Brown Corpus, 267–268
corpora in Natural Language Processing, 256
IC (information corpora), 268–269
data formats, 257
data sources, 257
document categorization, 274–277
Green Eggs and Ham, 256
in-database, 338–339
lemmatization, 258
morphological features, 266–267
NLP (Natural Language Processing), 256
parsing, 257
POS (part-of-speech) tagging, 258
raw text, collection, 260–263
search and retrieval, 257
sentiment analysis, 277–283
stemming, 258
stop words, 270–271
text mining, 257–258
TF (term frequency) of words, 265–266
DF, 271–272
IDF, 271–272
lemmatization, 271
stemming, 271
stop words, 270–271
TFIDF, 269–274
tokenization, 264
topic modeling, 267, 274
LDA (latent Dirichlet allocation), 274–275
web scraper, 262–263
word clouds, 284
Zipf’s Law, 265–266
text mining, 257
textual data files, 6
TF (term frequency) of words, 265–266
DF, 271–272
IDF, 271–272
lemmatization, 271
stemming, 271
stop words, 270–271
TFIDF, 269–274
TFIDF (Term Frequency-Inverse Document Frequency), 269–274, 285–286
time series analysis
ARIMA model, 236
ACF, 236–237
ARMA model, 241–244
autoregressive models, 238–239
building, 244–252
cautions, 252–253
constant variance, 250–251
evaluating, 244–252
fitted models, 249–250
forecasting, 251–252
moving average models, 239–241
normality, 250–251
PACF, 238–239
reasons to choose, 252–253
seasonal autoregressive integrated moving average model, 243–244
ARMAX (Autoregressive Moving Average with Exogenous inputs), 253
Box-Jenkins methodology, 235–236
cyclic components, 235
differencing, 241–242
fitted models, 249–250
GARCH (Generalized Autoregressive Conditionally Heteroscedastic), 253
Kalman filtering, 253
multivariate time series analysis, 253
random components, 235
seasonal autoregressive integrated moving average model, 243–244
seasonality, 235
spectral analysis, 253
stationary time series, 236
trends, 235
use cases, 234–235
white noise process, 239
tokenization in text analysis, 264
topic modeling in text analysis, 267, 274
LDA (latent Dirichlet allocation), 274–275
TP (true positives), confusion matrix, 224
TPR (true positive rate), 225
transaction data, 6
transaction reduction, Apriori algorithm and, 158
trends, time series analysis, 235
TRP (True Positive Rate), 185–187
ta() function, 245	two-sided hypothesis test, 105
type I errors, 109–110
type II errors, 109–110
typeof() function, 72

U
UNION ALL operator (SQL), 332–333
units of measure, k-means, 132–133
unstructured data, 6
Apache Hadoop, HDFS, 300–301
LinkedIn, 297
MapReduce, 298–299
natural language processing,
  18
use cases, 296–298
Watson (IBM), 297
Yahoo!, 297–298
unsupervised techniques. See clustering
users of data, 18

V
validation, association rules and, 157–158
variables
  categorical, 170–171
  continuous, discretization, 211
  correlated, 206
decision trees, 205
dependent, 162
factors, 77–78
independent, 162
input, 192
redundant, 206
VARIMA (Vector ARIMA), 253
vectors, R, 73–74
video footage, 16
  k-means and, 119
video surveillance, 16
visualization, 41–42. See also data visualization
exploratory data analysis, 82–85
  single variable, 88–91
grocery store example (Apriori), 152–157
volume, variety, velocity. See 3 Vs (volume, variety, velocity)

W
Watson (IBM), 297
web scraper, 262–263
white noise process, 239
Wilcoxon rank-sum test, 108–109
wilcox.test( ) function, 109
window functions (SQL), 343–347
word clouds, 284
work spaces, 10, 11. See also sandboxes
  Data preparation phase, 37–38
worker nodes, 301
write.csv( ) function, 70
write.csv2( ) function, 70
write.table( ) function, 70
WSS (Within Sum of Squares), 123–127

X-Z
XML (eXtensible Markup Language), 6
Yahoo!, 297–298
YARN (Yet Another Resource Negotiator), 305
Zipf’s Law, 265–266