Subject index

@ operator, 190
140 character limit, 370
1960s, 367

"A Boy Named Sue", 101
A/B testing, 5, 162
Aadam, 102
Aaron, Hank, 82
access
  databases, 490
  elements, 427
  variables, 447
accidental patterns, 149
accountability, 131
ACM, 143
actors, 382, 386, 387
actuarial tables, 49
add circles, 349
add rownames, 181
additional factor, 11
additional factors, 161
adjacency matrix, 395, 396
administration, database, 301
Adobe Illustrator, 254
Advanced R, 104, 438
adversarial system, 140
advertisers, 390
aesthetic, 33
  fill, 34
  graphical, 34, 387
  multiple, 34
Africa, 324
afternoon, 160
age, 196
aggregating, 70, 181, 275
airline delays, *see* flight delays
airline route maps, 345
airport, 150, 346
  codes, 282
  Reagan National, 285, 308
Akten, Memo, 28
Alaska, 287, 350
Albers projection, 324
Albert
  Jim, xviii, 88
  Réka, 380
Alderson, Sandy, 77
algorithmic thinking, 3, 441, 451
alias
  table, 283
alive, 49
Allaire, J.J., xviii
Alon, Noga, 382
alpha level, 166
Altavista, 390
alter database, 303
ALTER statement, *see* SQL
Amazon, 261, 401
  Web Services, 412
ambient temperature, 24
American Airlines, 79, 294
American League, 112
American National Standards Institute, 267
American Political Association, 144
American Statistical Association, 142, 143, 451
  statement on p-values, 167
American Statistician, 413
Amherst College, xvii, 330
Amherst, MA, 330
ampersand, 371
Anaheim Angels, 107
analogy, cooking, 14
analysis
  reproducible, 92, 454
  univariate, 24
analysis of ecological and environmental data task view, 436
analysis of pharmacokinetic data task view, 436
analysis of spatial data task view, 436
analytic (exact) result, 226
analytical skills, xv
analytics, 63, 413
  predictive, 8, 171, 173, 201, 205
anchoring, 358
AND operator, *see* SQL
Anderson, Wes, 30
anecdotal evidence, 24
Angelica, NY, 59
angle, 324
animation, 19
Anna, 58
annotated text, 52
ANSI standard, 267

Aotearoa (New Zealand), 421
Apache
  Drill, 412
  Hadoop, 409
  Spark, 409, 414
APIs, 117, 120, 127, 138, 141, 330, 360, 369
Apple R FAQ, 422
application programming interface, 117, 120, 127, 141, 330, 360, 369
apply family, 106, 349, 405
apportionment, 334
ArcGIS, 351
ArcView, 319
arguments, 434, 442, 443
array
  extract elements, 429
  indexing, 105
  rectangular, 93
  two-dimensional, 91
arrow, 52
arrows, sorting, 244
art, 28, 30
artificial neural network, 185
artistic talent, 28
arxiv.org, 360
ascending order, 70
aspect ratio, 92
assembly language, 404
assertions, 451
assignment operators, 426
Association for Computing Machinery (ACM), 143
associations, 159, 463
assumption, normality, 474
astronomy, 363
athletes, 28
Atlanta, GA, 208, 346
Atlantic 10 conference, 394
attack ad, 10
attributes, 34, 93
  R, 431
attribution, 24
Auckland, University of, 421
audience, intended, xvi
audits, data, 141
austerity measures, 135
authorship, 24, 367
automation, 104
average
  running, 444
avoiding for loops, 106
axis, 38
axis labels, 26
axis scaling, 12
aye votes, 216

baby names, 7, 48, 56, 93, 101, 244, 309, 410, 495
backticks, 190
Bacon, Kevin, 382
bacteria, 329
Bale, Christian, 389
ballot, 95, 212
bank tellers, 231
Banquo, 359
bar graph, 10, 20, 41
Bar–Noy, Amotz, 382
Barabási, Albert-László, 380
barrels of oil, 33
base graphics, 33
baseball, 7, 72, 82, 88, 104, 107, 402
  analytics, 6
  camera-tracking data, 402
  offensive statistics, 82
  play-by-play data, 402
  rules, 6
basketball, 391, 392
bathrooms, 120
Batman, 101
Batman: The Dark Knight Rises, 383, 389
batter value, 6
batting, 82
Bayes
  naïve, 183–186, 189, 191, 200
  Theorem, 183
Bayesian inference task view, 436
Bayesian statistics, 236
Bears, Chicago, 28
Beatles, 244, 246, 248, 254, 367–369
bedrooms, 120
beginning of line, 358
bell-shaped, 155
Bella, 48
Belmont Report, 143
benchmarking, 175
Bentley College, 140
Berkeley Initiative for Transparency in the Social Sciences, 144
best job, xv
BETWEEN operator, *see* SQL
betweenness centrality, 378

Beveridge, Andrew, 212
bias, 166
  publication, 143
  selection, 464
  variance tradeoff, 192, 471
bichromatic choropleth map, 340
big data, 3, 149, 261, 401, 413
big memory, 403
BigQuery, 410, 412
bike trail, 464
billing, top, 383
bills of mortality, 3
Bills, Buffalo, 19
binary, 404
binwidth, 41
Bioconductor, 435
biologists, 205
birds (Ordway), 120, 123
births, 93
Bivand, Roger, 352
blindness, color, 18, 19
blocks of code, 455
blood pressure, 98, 99
BMI, 46, 196, 475
body mass index, 475
Boehner, John, 13
bold font, 455
Bonferroni correction, 166
book website, xvii
Boolean
  operations, 428
Boolean operators, 358
Boolean pairs, 28
bootstrap, 115, 155
  decision trees, 181
  distribution, 157
Bostock, Mike, 243
Boston Red Sox, 83
Boston, MA, 208
botulism, 131
bounding box, 327
box, bounding, 327
boxplot, 44, 160
Boynton, Nancy, xviii
BP_narrow, 99
BP_wide, 98, 99
Bradley Airport, 275
Brain, Behavior, and Immunity, 142
branches, 179, 205
breach, ethical, 135
break down problems, 441
breast cancer, 48
Brewer, Cynthia, 19, 30
Bridges of Königsberg, 377
Brin, Sergey, 390
British invasion, 367
Broad Street pump, 318
Broadwick (formerly Broad) Street, 322
Brooklyn Dodgers, 6
Brooklyn, NY, 331
Brown, Scott, 14
browser, 118, 243
  debugging, 450
Bryan, Jenny, 457
budgeting time, 27
Buffalo Bills, 19
bulk updates, 309
Bureau of Labor Statistics, 227
Bureau of Transportation Statistics, 79, 418

C++, 404
calculate running average, 444
calculate score, 366
California Angels, 107
calling functions, 433
CamelCase, 104
cancelled flights, 290
cancer, 222
cancer genetics, 48
canonical graphics, 46
capital gains, 175, 176
capitalization
  in object names, 103
captions, 26
Caris, Jon, xviii
cars, 206
CART, 173
Cartesian coordinates, 18
Cartesian plane, 41
cascading delays, 159
case, 93, 97
  in codebook, 98
  sensitivity, 423
  studies, xv
  unique, 97
cases, special, 441
Cash, Johnny, 101
categorical, 15
  outcome, 475
  predictor, 467
  scale, 18

variable, 53, 96
causal inference, 167
causality, 44
causation, 162, 173
cell lines, 48
Census Bureau, 3, 174, 334, 352
Census tracts, 22
Center for Open Science, 144
centrality, 377, 378, 390
CEO, 134, 137
cesspit, 329
Çetinkaya–Rundel, Mine, xviii
Chadwick, Henry, 6, 402
chaining operations, 66, 73
Challenger disaster, 23
CHANCE magazine, 144
change database, 303
changes, tracking, 457
changing font size, 455
changing graphic size, 455
Channing Laboratory, 142
character data, 355
character limit, 370
character sets, 358
character string, 125, 355
character vector, 447
charges, medical, 39
Che, Jonathan, xviii
cheat sheets, 58, 88, 258, 457
checkboxes, 248
chefs, great, 14
chemometrics and computational physics task view, 436
Cherry Blossom Race, 97
Chestnut Street, 464
Chicago Bears, 28
Chicago White Sox, 83
Chicago, IL, 208
chicken pox, 131
child abuse, 131
cholera, 131, 317, 328
choosing names, 48
choropleth map, 340, 343
Church, Angelica Schuyler, 59
Cincinnati, OH, 346
circles, 349
Citi Field, 331
Citizens United, 9
civil rights, 140
class label, 173
class methods, 431
classic books, 374
classification, 201
classification and regression trees (CART), *see* CART
classification rate, 192
classifiers, 173, 183, 186
perfect, 191
cleaning data, 116, 120, 425
Cleveland Indians, 83
Cleveland, OH, 278
click-and-drag, 139
client, 137, 139
client–server, 485
client-server model, 267
client-side Web applications, 243
climate change, 133
clinical trial design, monitoring, and analysis task view, 436
clinical trial registries, 144
clinical trials, 166, 167
closure, triadic, 380
cloud storage, 412
cluster, 314
hierarchical, 206, 208
cluster analysis and finite mixture models task view, 436
clustering, 205, 212, 215
clusters, Spark, 411
code
airport, 282
blocks, 455
completion, 421
control systems, 454
defensive, 450
ethical, 143
examples, xvii
review, 142
solutions, 441
codebook, 97, 120
coding errors, 135
coefficient
of determination, *see* R^2
regression, 474
true, 472
coercion, 428, 447
data frames into matrices, 431
matrices into data frames, 431
coffee, 162
cognitive overload, xv
cognitive skills, 149
collaboration, 457

collapse, 70, 275
college
  basketball, 391
  entry, 162
colleges, liberal arts, xvii
colon cancer, 48
color, 10, 19, 56, 251
  blindness, 18, 19
  diverging palette, 341
  gradient, 477
  scheme, 340
ColorBrewer, 19, 386
Columbia University, 143, 166
column names, 64
columns, 93, 412
combining tables, 79
comma separated, 265, 303
command history
  R, 422
command line, 486, 489
commands, 441
comments, 429
common sense, 134
Commonwealth Games, 28
communication skills, xv
communications, privileged, 140
commuting routes, 331
compare versions, 457
comparison operators, 427
competing interests, 143
compilation, 404
complexity parameter, 179
Comprehensive R archive network, 422
compromise, 404
computational linguistics, 355
computational thinking, xv
computer programming, 3
computer science, 4, 390
computing, 3
  parallel, 404
  scriptable, 454
  statistical, 221, 236
CONCAT function, *see* SQL
concept index, xvi
conclusions, 166
conditional
  inference, 201, 221
  probability, 183, 236
  regression parameter, 468, 479
conference, Atlantic 10, 394
confidence band, 471
confidence interval, 115, 153, 157, 161, 441, 472
confidentiality, 134
conflict of interest, 137
conflicts, 431, 437
confounding, 161–163
confusion matrix, 180, 189
congressional districts, 333, 336, 340, 351
connected component, 389
connected graph, 377
consensus statement, 131
consent, 138
Conservative and Unionist Party, 215
conservative, socially, 216
consistency, internal, 136
Consolidated Standards of Reporting Trials (CONSORT), 144, 167
constraints, 201
context, xvi, 15, 19, 27
continents, 211
continue execution, 450
contributions, 9
control structures, 423
control, version, 454, 457
convenience, 404
conventions, naming, 103
convergence, 234, 444
convert
  character string to date, 123
  character to numeric, 122
  tall to wide, 100
  types, 450
  wide to tall, 100
cooking analogy, 14
Coolidge Bridge, 331
coordinate reference system (CRS), 325
coordinate systems, 15, 18
  radial, 22
coordinates, 322
copyrights expired, 374
cores, 405
corpus, 355, 364, 408
correct algorithm, 441
correction for multiple comparisons, 144, 223
correlation, 189, 367, 465
  does not imply causation, 162
counting patterns, 356
countries, 33
courses, swirl, 425
courts, 138, 140

CRAN (Comprehensive R Archive Network), 422, 438
CRAN task views, *see* task views
Cranston, Bryan, 386
create databases, 301, 302
CREATE statement, *see* SQL
create table, 302
create temporary file, 372
creating presentations, 27
credits, 383
critical analysis, 15
cross join, 281
cross-validation, 183, 188
crossings, trail, 464
CRS, 325
CSV file, 116, 265
CUDA, 406
cues, visual, 15, 33, 34, 56
cumulative sum, 445
curated guide to learning R, 425
current population, 93
customers, 231
customizing graphics, 33
Cygwin, 485

D3.js, 243, 254
damage, O-ring, 24
damages, 136
damping, 397
data
  analysis, 171, 453
  arrange, 107
  art, 28, 30
  big, 261, 401, 413
  cleaning, 116, 120
  definition, 149
  demographic, 352
  disclosure, 140
  documentation, 97
  encoding, 15
  extracting meaning, 4, 149, 413
  fabrication, 143
  frames, 63, 93, 262, 426, 429–431
  friendly formats, 116
  gathering, 101
  generation, 236
  graphics, 9, 33
  habits of mind, 88
  idioms, 63
  indexing, 105
  ingesting, 120
  intake, 116
  integration, 318
  interpretation, 149
  journalist, 88
  medium, 261, 403
  meta, 142, 327
  microarray, 222
  mining, 3
  missing, 167, 464
  network, 5, 48, 377
  observational, 162, 165
  omitted, 24
  pipeline, 417
  provenance, 116
  reading from a URL, 118
  real-world, 221
  release, 402
  safeguards, 140
  scraping, 116, 135
  sets, xvii, 7, 417
  sharing, 143
  spatial, 5, 317, 352
  spreading, 100
  storage, 141
  structure, 317
  table, 93
  tables, 93, 244
  technologies, xv
  test, 175, 188
  text, 5, 317, 355
  tidy, 88
  to ink ratio, 41
  to understanding, 221
  training, 175, 188
  transforming, 93
  verbs, 63
  viewer, 421
  visualization, 3
  weather, 7
  wide and narrow, 98
  wrangling, xv, 3, 63, 88, 94, 136, 447
  wrangling cheat sheet, 88
data science, xvi, 4, 8, 221, 318, 401
  ethics, 143
  papers, 360
Data Science Association, 143
Data Science Journal, 363
data scientist, xv
data structures, 426
data verbs
  arrange, 99

- gather, 100
- group by, 101
- mutate, 95
- spread, 100
- summarize, 101

database, 261
- access, R, 490
- administration, 301
- baseball, 72, 402
- design, 285
- Internet movie, 382
- keys, 82, 303
- load, 495
- MySQL, 485
- PostgreSQL, 485
- real estate, 141
- relational, 98, 117, 266
- schema, 303, 310, 495
- server, 267, 485
- SQL, 485
- SQLite, 485
- tweets, 372

date variables, 67, 123, 127
dates and times, 272
datums, 352
dbf files, 320
DBIConnection, 409
De Veaux, Richard, 120
Death of a Salesman, 28
debt crisis, 135
debugging, 450
deception, 221
decimal places, 303
decision tree, 173
decisions, xv, 166
decomposition, singular value, 213
deep analytical skills, xv
default reference category, 467
default value, 442, 443
defensive coding, 450
definition
- big data, 401
- data science, 5, 8
- statistics, 5, 152

degree centrality, 378
degree distribution, 380
deidentified data, 135
delays, airline, *see* flight delays
delete objects, 427
Delta Airlines, 159, 345, 347
Democratic, 10, 335
demographic data, 352
dendrogram, 205, 208
density curves, 12
Department of Energy, 206
departure time, 159
dependency management, 438
dependent variable, 463
depression, 142
descending order, 70
describe database, 302
DESCRIBE statement, *see* SQL
design of experiments (doe) and analysis of experimental data task view, 436
design of simulations, 233
designated hitter, 113
detach packages, 437
Detroit, MI, 346
diabetes, 46, 172, 196, 475
Diaconis, Persi, 382
diagnostics, residual, 474
dialects, SQL, 267
diameter, 377
differential equations task view, 436
Digital Ocean, 412
dimension reduction, 211
directed edges, 377
directed graph, 185, 390
direction, 21
disclosure avoidance, 140
discrimination, employment, 134, 138
disease, 196, 317
disk, hard, 265
displaying objects, 437
dissecting graphics, 20
distance, 207, 331
distance measure, 377
distributed architectures, 314
distributed computing, 404
distributed file system, 409
distribution
- degree, 380
- donations, 12
- exponential, 232
- normal, 155
- number of characters, 370
- Poisson, 232
- probability, 225
- sampling, 152
- stationary, 391
- uniform, 397

districts, congressional, 333
diverging palette, 20, 341
DNA, 172
document term matrices, 365
documentation
  codebook, 97
  R, 423
documentation for R, 438
Dodgers
  Brooklyn, 6
  Los Angeles, 83
domain knowledge, 5, 6
downloading
  code examples, xvii
  tweets, 370
dplyr package, 3
  issues on GitHub, 407
Dremel, 412
DROP TABLE statement, *see* SQL
drug company, 135
dummy variables, 467
Duncan, 359
duplicates, 155
Duquette, Jim, 77
dygraphs, 246
dynamic maps, 332
dynamic Web applications, 248, 421

earth, 208
eating hot dogs, 254
eBay, 28
eccentricity, 377, 390
ecological disruption, 134
econometrics task view, 436
economic indicator, 227
economic productivity, 33
economically conservative, 216
economics, 135
edge, 48, 233, 346, 377, 384, 391
  centrality, 378
  effect, 230
  multiple, 396
  weight, 392
effect size, 167
effective decisions, xv
effective presentations, 23, 27
efficient, 305
  algorithms, 441
  databases, 301
  SQL, 280
Efron, Bradley, 167
eigenvector centrality, 378, 390
eikosogram, 46
election
  ballot, 95
  campaign, 10, 333
  federal, 9
  Minneapolis mayoral, 94
  rank choice, 97
electronic computers, 3
electronic repository, 360
ellipsoids, 352
Elmer, 49
else statement, 423
Elynor, 93
embarassingly parallel, 404
empirical finance task view, 436
empirical research, 457
employee selection, 134, 138
employer, 137
employment discrimination, 134, 138, 139
encoding data, 15
encoding variables, 120
end of line, 355, 358
Energy Department, 206
ensemble methods, 186
entropy, 174
environment, 437, 446
ephemeral table, 288
epidemiology, 167, 317
Epiinfo files, 117
EPSG, 326
epsilon, 463
equal area projection, 324
equal variance, 474
Erdős
  number, 382
  Paul, 378
Erdős–Rényi random graphs, 379
error matrix, 180, 189
error rate, 166
errors, 463
  coding, 135
  minimizing, 450
  prediction, 189
Esri, 319
estimators, 4
ethical principles, 131, 134, 136, 137, 162
ETL, 309
etl package suite, 417
Ettinger, Caroline, 212
Euclidean distance, 207

Euler, Leonhard, 377
Europe, 287
European
  debt crisis, 135
  Petroleum Survey Group (EPSG), 326
evaluate expression, 450
evaluating models, 188
evaluation, lazy, 451
evaluation, non-standard, 448
evening, 160
evidence, 167
evolutionary
  biologists, 205
  tree, 172
exabytes, 401, 402
exact result, 226
exam, high-stakes, *see* SAT scores
example code, downloading, xvii
Excel, 91, 117, 139, 206, 265, 402
Excite, 390
execution, 450
exercises, xvi
expected, 444
  flight delay, 159
expected winning percentage, 73, 110
expert scrutiny, 137
expert witness, 140
expired copyrights, 374
EXPLAIN statement, *see* SQL
explaining to humans, 454
explanatory variables, 173, 463
exploratory data analysis, 171
exponent, 110
exponential random variable, 232
expressions, 426, 441, 448, 450
  regular, 193, 253, 254, 355, 374
ExpressJet, 294
extended case studies, xv
extract from objects, 429
extract information, xv, 149
extract number, 127
extract values, 366
extract, transform, load, 309, 417
extracting meaning from data, 4, 413
extrapolation, 24, 468
extreme value analysis task view, 436

FAA, 282
fabrication of data, 143
Facebook, 261, 377, 401
facet, 19, 43, 53
facets, 38
factors, 125, 467
  drop levels, 46
Fahrenheit temperature, 464
fail to reject null, 166
false discovery rate, 223
false findings, 453
false positive, 191
falsehoods, 131
FAQ
  Apple R, 422
  R, 425
  Windows R, 422
FDA, 135
features, 210
FEC, 9
fecal bacteria, 329
Federal Election Commission (FEC), 7, 9, 334
federal elections, 9
female names, 53
Feynman, Richard, 23
field station, 418
file
  list, 310
file, source, 454
file-drawer problem, 143
files
  loading, 116
  readable, 454
  saving, 116
fill
  aesthetic, 34
  scale, 477
finding packages, 435
finish execution, 450
firearm murders, 132
first base, 6
fitted values, 474
five idioms, 63
FiveThirtyEight.com, 49, 56, 253, 289
flat-file database, 265
flight delays, 7, 79, 150, 159, 261, 275, 289, 302, 305, 418
Florence, MA, 464
Florida legislature, 132
flowchart, 173
Flushing, NY, 331
font size, changing, 455
fonts, 253
Food and Drug Administration (FDA), 135

food handling, 228
foot length, 46
for loops, 104, 423
Ford, Gerald, 67
foreign key, 303
FOREIGN KEY definition, *see* SQL
forest, random, 181
forking paths, garden of, 166, 454
format
  data table, 116
  date variables, 123
  PDF, 454
formatted output, 455
forms, normal, 285
formula, 172, 199, 464
found data, 162, 165
Foundation for Statistical Computing, R, 421
framework for graphics, 11, 14
Frenett, Patrick, xviii
frequency inverse document, 365
frequently asked questions, *see* FAQ
Friday, first, 227
fuel economy, 206, 210
Fukushima Daiichi, 127
full text, 374
function, 172, 247, 441, 448
  call, 450
  calling, 433
  examples, 425
  name conflicts, 437
functional form, 470
funding network, 13
fundraising, 9

Gall–Peters projection, 324
Gapminder, 91
garden of forking paths, 166, 454
gastro-intestinal events, 135
gather data, 98, 101
GDP, 33
Gelman, Andrew, 30, 144, 166
gender, 53, 93
gender neutral names, 101
generate random numbers, 235
generic functions, 417
genetics, cancer, 48, 222
Gentleman, Robert, 421
geocoding, 330
geodesic, 377
geographic coordinates, 18
geographic information, 140
geographic positioning systems, 326
geography, 48, 317
geolocated, 371
geom, 34, 50
  with new data, 38
geometry, 226
George Washington University, 392, 394, 396
geospatial coordinates, 324
gerrymandering, 343, 352
getting and cleaning data, 425
ggplot2 package, 3
  themes, 250
ggplot2 plotting commands, 46
ggvis, 246, 258
Gibb, Robin, 101
Gilman, Scott, xviii
Gini coefficient, 176
GitHub, 409, 420, 421, 435, 451, 454, 457
  install from, 434
  issues, 407
  learning, 457
global environment, 446, 447
global temperature, 133
glyph, 33, 56
  changing, 34
goodness of fit, 465
Google, 261, 330, 401
  BigQuery, 410, 412, 414
  Cloud Storage, 412
  Drive, 454
  Earth, 326, 351
  Maps, 326, 351
  n-grams, 374
  page rank, 378, 390, 391
  R style guide, 104, 431
  scale, 414
  sheets, 454
  spreadsheet, 91, 117, 265
Gordon–Levitt, Joseph, 389
Gorman, Michele, 28
governor, 140
GPS systems, 326
GPU computing, 406
grades, 228
gradient, scale, 477
grammar
  data wrangling, 63
  of graphics, 30, 33, 58
  regular expressions, 355

Grand Central Parkway, 331
graph, 233
  random, 378
  theory, 13, 233, 377
  visualization, 385
graphic displays and dynamic graphics and graphic devices and visualization task view, 436
graphic size, changing, 455
graphical framework, 11
graphical models in R task view, 436
graphical perception, 18
graphical ploys, 131
graphical processing unit, 406
graphical themes, 250
graphical user interface
  RStudio, 421
graphics
  aesthetic, 33, 34
  annotated text, 52
  arrow, 52
  bar graph, 41
  boxplot, 44, 160
  cheat sheets, 58
  eikosogram, 46
  grammar, 30, 33, 58
  grid of plots, 38
  interactive, 243
  layers, 38
  legend, 38
  mosaicplot, 46
  multivariate, 43
  pie charts, 41
  scale, 18, 37
  scatterplot, 43
  taxonomy, 46
  title, 52
  univariate, 39
graphs
  random, 233
  versus tables, 26
Graunt, John, 3
great chefs, 14
greedy strategies, 173
Greenland, 324
grid computing, 404
grid of plots, 38
gross domestic product (GDP), 33
GROUP BY clause, *see* SQL
grouping, 70
groups, iteration over, 110
guide to packages, 435
guidelines
  design of simulations, 233
  ethical conduct, 143
  for data science programs, 414
  for statistics programs, 451
guides, 38
gun control, 132
Gutenberg Project, 355, 374

hacking, 141
Hadoop, 409
Hadron Collider, 402
Hamilton, Alexander, 59
handling and analyzing spatio-temporal data task view, 436
Hanna, Maya, xviii
Hansen, Mark, 28
hard disk, 265
Hardin, Johanna, xviii
Hardy, Tom, 389
Harrell, Frank, 434
Harrison, George, *see* Beatles
Harvard University, 142, 144
harvesting, 141
hashtag, 373
Hastie, Trevor, 167
Haverford College, 457
HAVING clause, *see* SQL
Hawaii, 287
head node, 314
health care, 172
Health Insurance Portability and Accountability Act (HIPAA), 140
health records, 140
health study, 470
health violations, 7, 228
heart attacks, 135
heat maps, 18
heat type, 120
heavy tail, 474
hedging bets, 188
height, 44, 471
HELP study, 21, 41, 98
help system, 423
  R, 423
  R packages, 438
hemisphere, 172
Herndon, Thomas, 135
Hesterberg, Tim, 167
heteroscedasticity, 474

Hewlett Packard, 267
hidden layer, 185
hierarchical clustering, 206, 208
high-earner, 174
high-performance and parallel computing with R
  task view, 414
high-performance and parallel computing with R task view, 436
high-stakes exam, *see* SAT scores
HIPAA, 140
Hippocratic Oath, 131
histogram, 11, 41
history
  of commands, 422
  R, 421
  version, 457
HIV, 91
Hive, 409
Hoaglin, David, 479
hold-out sample, 175
Hollerith, Herman, 3
Hollywood, 382, 385, 389
home run, 6, 82, 85
homeless, 41
homogeneous subsets, 173
homoscedasticity, 474
honesty, 137
Horton, John, xviii
hospital, 317
hospitalization, 140
hot dog eating, 254
hour, minute, second format, 123
House of Representatives, 7, 12
house prices, 120, 141
hovering, 247
how to lie with statistics, 131
Howe, Bill, 406
HTML
  files, 117, 332
  format, 367, 454
  tables, 118
htmlwidgets, 243, 258, 332
Huff, Darrell, 131
humidity, 411
Hunt's algorithm, 174
hybrids, 207
hypothesis, 172, 221, 472
  null, 166, 223
hypothesis testing, 165
IBM Corporation, 3
ice cubes, 23
idioms for data, 63
if statement, 423
IGNORE clause, *see* SQL
Ihaka, Ross, 421
imaging, medical, 190
IMDb, 382, 383
immune response, 142
implementation, SQL, 485
importance, 181
  of captions, 26
IN operator, *see* SQL
in operator, 91
in statement, 423
inches, 469
income, 174
independence, 473
independent variable, 463
index, 273, 356
indexing, 105, 449
  lists, 428
  vector, 427
indicator variables, 467
indicator, economic, 227
indices, 303, 304
  R, xvi
  subject, xvi
infarction, myocardial, 135
inference, 149, 463, 472
  bootstrap, 155
  causal, 167
  conditional, 221
  modern, 167
  perils, 165
  trees, 201
inflation, 174
influential point, 474
information gain, 174
ingesting data, 116, 120, 369
ingesting text, 367
inner join, 281
insert data, 491
INSERT statement, *see* SQL
insights from outliers, 158
inspections, restaurant, 228
installing
  packages, 434
  R, 422
  RStudio, 422
  SQL server, 301

Institute for Quantitative Social Science, 144
instructing computers, 454
insurance records, 140
integer codes, 121
integrity, 144
  ethical, 131, 457
  referential, 82, 304
intellectual traditions, 4
intended audience, xvi
interaction model, 470
interactive
  courses, 425
  graphics, 243
  maps, 332
  tables, 244
  Web applications, 247
intercept, 463
internal consistency, 136
Internet movie database, 382
Internet use, 33
interpretation, 468
interpreted languages, 404
interpreting data, 149
intervals
  confidence, 115, 153, 441
  presidential terms, 66
introduction
  R, 421, 423
  RStudio, 421
Ioannidis, John, 453
irreproducibility, 454
issues, tracking, 457
iteration, 104, 113
  over subgroups, 110
iterative process, xvi

James, Bill, 6, 110
Japanese nuclear reactors, 126
Javaid, Azka, xviii
JavaScript, 243, 244, 254, 332
Jessie, 53
Jets, New York, 19
jitter points, 476
job title, 139
job, best, xv
jobs report, 227
Johnson, Lyndon, 67
join
  left, 285
  tables, 79, 85, 87, 98, 121, 142, 281, 307, 338
JOIN clause, *see* SQL
Jordan, Michael, xv, 4
Joseph, 49, 50
Josephine, 53
journalism, 88, 134
JSON
  files, 117
  objects, 370, 407, 412
Julia, 413
jurisdiction, 131
juvenile pleasures, 345

k-fold cross-validation, 188
k-means, 210
k-nearest neighbor, 182, 200
Königsberg bridges, 377
Kadir, 93
Kaggle, 391
Kennedy, John, 67
Kentucky University, 392, 394
kernel smoother, 41
Kevin Bacon number, 382, 389
key, 82, 303, 369
  primary, 303
key-value pairs, 412
Keyhole Markup Language (KML), 351
Kim, Albert, xviii, 141
Kim, Eunice, xviii
Kimmel, John, xviii
King of Scotland, 359
Kleinman, Ken, xviii
KML, 351
knitr, 454
knowledge
  domain, 6
Knuth, Donald, 454
Kusiak, Caroline, xviii

La Salle, 397
lab notebooks, 453
label, class, 173
Labour Party, 215, 217
lack of fit, 180
Lady Macbeth, 359
Lahman, Sean, 88
Laird, Nan, 382
Lambert projection, 324
Lancet, 135
language processing, 355

language, structured query, 261
languages
  interpreted, 404
Large Hadron Collider, 402
large networks, 377
large numbers, law of, 444
laser sensor, 464
LASSO, 192, 201
latitude, 210, 317, 322, 373
lattice graphics, 33
law of large numbers, 444
law, stand your ground, 132
lawyer, 134
layer, 43
layers, 19, 38, 185, 320, 332
lazy evaluation, 451
lazy learners, 182
leaflet, 244, 332, 343, 347
learning from data, xv
learning R, 425
learning, statistical, 171, 201, 205
Least Absolute Shrinkage and Selection Operator (LASSO), 201
least squares, 463
left join, 281, 285
LEFT JOIN clause, *see* SQL
legal negotiations, 136, 139
legal system, 137, 140
legends, 38
legislature, Florida, 132
length of office, 66
Lennon, John, *see* Beatles
less volume, more creativity, xv
lethal force, 132
levels, 96
leverage, 474
Lexington Street, 322
Li, Priscilla, xviii
liberal arts colleges, xvii
Liberal Democrats, 215
libraries, 404
library
  help, 438
  R, 434
life tables, 49
LIMIT clause, *see* SQL
limitation, methodological, 137
linear algebra, 213, 390
linear model, 162
linear regression, *see* regression
linear relationship, 465
linearity, 473
linguistics, 355
links, 390
Linux, 485
  installation, 422
list files, 206, 310
list object, in R, 119
list of data sets, 417
lists, 118, 428
  extract elements, 429
literate programming, 454
literature, 28
LOAD DATA statement, *see* SQL
load files, 116
load into SQlite, 495
loading data, 309
locally optimal strategies, 173
locations, 331
log scale, 176
logic of science, 453
logical operator, 427
logical vectors, 360
logistic regression, 162, 173, 193, 475
London, England, 3, 317
long format, 92
Long Island Expressway, 331
longitude, 210, 317, 322, 373
loops, 104, 106
Los Angeles Angels, 107
Los Angeles Dodgers, 83
Los Angeles, CA, 208
low-rank, 212
lung cancer, 48, 162, 222
lurking variables, 163
Lycox, 390
lying, 136

Mac OS X, 485
Macalester College, 212
Macbeth, 355
MacDuff, 356
machine learning, 3, 171, 190, 205, 218, 402
  library, 411
machine learning and statistical learning task view, 218, 436
Macintosh R FAQ, 422
MacLeish field station, 7, 44, 411, 418
Maclester College, xvii
Madness, March, 391
MAE, 189

Major League Baseball, 72, 82
majority rule, 181, 182
making up data, 221
male names, 53
malware, 141
managing data, 94
managing packages, 457
MapReduce, 406, 408, 409
maps, 18, 48, 317, 322
  airline route, 345
  choropleth, 340
  dynamic, 332
  projections, 352
  proportional symbol, 344
  road, 208
March Madness, 391
Marchi, Max, 88
margin of victory, 392
mark, 33
markdown, xvi, 142, 244, 454
  cheat sheets, 457
  output format, 456
  templates, 455
Markov Chain Monte Carlo, 236
Martin, Paul, 217
masking, 431, 437
mass, movable, 391
Massachusetts, 13, 22
  governor, 140
mastering data science, 417
matching characters, 357
math SAT scores, 20, 41, 43
mathematical statistics, 167
mathematics, 5, 413
MathSciNet, 382
matrix, 173, 212
  adjacency, 395, 396
  algebra, 213
  confusion, 180, 189
  distance, 208
  document term, 365
  extract elements, 429
  indexing, 429
  multiplication, 397, 428
  R, 429
  sparse, 365
  transition, 391
  voting, 218
maximum delay, 150
mayoral election, 94
McCartney, Paul, *see* Beatles
McIntyre, Mike, 335
McKinsey & Company, xv
McNamara, Amelia, xviii
mdsr package, 417
mean, 152, 153
mean absolute error (MAE), 189
measure of evidence, 167
measure, distance, 377
mechanism, 221
medical image analysis task view, 436
medical imaging, 190
medical procedures, 38
medical records, 140
Medicare, 38
medium data, 261, 403, 417
meeting times, 225
melanoma, 48
memory (RAM), 265
memory, big, 403
Mercator projection, 324, 350
Merck, 135
merge tables, 142, 307, 338
mergers, airline, 292
merging tables, 79, 281
Mescon, Cory, xviii
meta-analysis task view, 436
metacharacter, 357
metadata, 142, 327, 360, 431
method overloading, 68
methodological limitation, 137
methodology, statistical, 149
methods, 431
Mets, New York, 72
Mexico, 287
microarray, 222
Microsoft, 267
Microsoft Excel, 91, 117, 139, 206, 402
Microsoft Word, 454
  format, 455, 456
Middlebury College, 141
miles, 208
miles per gallon, 206
Milgram, Stanley, 382
Miller, Arthur, 28
Minaya, Omar, 77
minimizing errors, 450
minimum, 103
mining text, 28, 141, 356
Minneapolis mayoral election, 94
Minneapolis–St. Paul, 279, 346
misclassification rate, 192

misleading graphics, 131, 132
missing data, 105, 167, 464
missing values, 82, 87
mistakes, 143
MIT, 140
mode of storage, 431
model
  assumptions, 473
  ensemble, 186
  evaluation, 188
  formula, 172
  interaction, 470
  linear, 162
  logistic, 162, 173, 193, 475
  non-linear, 470
  null, 175, 180, 193, 223, 465
  overfitting, 179, 188
  predictive, 172
  regression, 44, 161, 163, 171, 463
  regularization, 201
  statistical, 159, 329
  to generate data, 221
  visualization, 198
modern inference, 167
modularity, 234
Moneyball, 6
MongoDB, 267, 412
monitor, MySQL, 487
Montreal, 287
Monty Hall problem, 236
more creativity, less volume, xv
morning, 160, 275
mortality, 3
Morton Thiokol, 23
mosaicplot, 46
Mosteller, Fred, 382
mouse-over, 247
movable mass, 391
movement of athletes, 28
movies, 30, 382
multidimensional data, 63
multiple comparisons, 144, 166, 223
multiple edges, 396
multiple regression, 411, 463, 468
multiple tables, 98
multiples, small, 19
multiplication, matrix, 397
multivariate displays, 43
multivariate statistics task view, 436
murders, 132
myocardial infarction, 135
MySQL, 382, 409, 411, 485
  monitor, 487
  server, 262
  syntax, 345

n-gram, 374
naïve Bayes, 183–186, 189, 191, 200
name conflicts, 431, 437
named arguments, 434, 443
named lists, 428
names
  baby, 7, 48, 56, 93, 244, 309, 410
  gender neutral, 101
  special characters, 127
naming conventions, 103
naproxen, 135, 139
Narraido, Tasheena, xviii
narrow data, 98, 100
NASA, 23
National Academy of Sciences, 143
National Cancer Institute, 48
National Health and Nutrition Examination Study (NHANES), *see* NHANES
National Party, 215, 217
National Review, 133
native file format, 116
natural language processing, 355
natural language processing task view, 436
nay votes, 216
NCAA, 391
nearest neighbor, 182
needle in haystack, 356
negative results, 142
negligence, 143
negotiations, legal, 136
neighbor, nearest, 182
nesting, 73
nesting function calls, 443
network science, 3, 5, 48, 185, 233, 377
  history, 378
networks, 13
neural network, 185
neutral assessment, 134
New England Journal of Medicine, 135, 139
New Haven, 374
New Jersey, 38, 39
new users
  R, 423
New York City, 150, 160, 228
New York Jets, 19

New York Mets, 72, 76, 331
New York Times, 14, 227, 243, 254
New York Yankees, 112
New York, NY, 346
New Zealand (Aotearoa), 421
Newark, NJ, 162
next statement, 423, 450
NHANES study, 44, 172, 196, 470
Nixon, Richard, 67
node, 185
nodes, 118, 233
Nolan, Deborah, 30
non-farm payroll, 227
non-linear models, 470
non-standard evaluation, 448, 451
nondisclosure, 140
normal distribution, 155
normal forms, 285
normal random variables, 227
normality assumption, 474
normality of residuals, 474
normalization, 344
North Carolina, 333, 334, 340, 351
Northampton, MA, 172, 330
Northwest Airlines, 345
NoSQL, 412
notebooks, lab, 453
nouns, 28
NP-complete, 173
NSE (non-standard evaluation), 448
nuclear reactors, 126
null hypothesis, 221, 472
null model, 175, 180, 193, 223, 465
number of characters, 370
number, parsing, 127
numeric scale, 18
numeric vector, 442
numerical, 15
numerical mathematics task view, 436
NumPy, 413
nutrition study, 470
Nuzzo, Regina, 167
NVIDIA, 406
NYC Open Data, 127

O(n), 305
O-rings, 23
Oakland A's, 6
oath, Hippocratic, 131
Obama, Barack, 10, 333, 378
object-oriented programming, 190, 321, 431
objects
  coercion, 428
  displaying, 437
  list, 119
  R, 426
  reactive, 248
  remove, 427
  rules for naming, 103
oblate spheroid, 324
obligations, 131
observation, 93, 97
observational data, 165, 167
OBSP, 85
ocean levels, 134
Octave files, 116
Odum Institute for Research in Social Science, 142
OFCCP, 134, 138
offensive statistics, 82
Office of Civil Rights, 140
Office of Federal Contract Compliance Programs (OFCCP), 134, 138
official statistics and survey methodology task view, 436
Ohio, 13
oil, number of barrels, 33
OkCupid, 135, 138, 141
OLAP, 63
Omegahat, 435
omitted data, 24
on base slugging percentage (OBSP), 85
online library, 374
Open Data NYC, 127
Open Intro Statistics, 479
Open Science, 144
Open Street Maps, 326
open-source, 409
OpenCL, 406
OpenIntro Statistics, 152
OpenPsych Forum, 135
OpenStreetMaps, 244
operations research, 190
operations, vector, 104
optimization and mathematical programming task view, 436
option file, 488
optional argument, 442
options in R, 433
OR operator, *see* SQL
or operator, 427
Oracle, 267

oracle, Kristen Stewart, 387
ORDER BY clause, *see* SQL
order, descending, 70
ordinal variable, 69
Ordnance Survey, 326
Ordway birds, 120
other factors, 161, 162
Ott, Mel, 6
outcome variable, 463
    measured, 172
    unmeasured, 172
outer join, 281
outliers, 157, 474
output
    file, 454, 455
    format, 456
    formatted, 455
ovarian cancer, 48, 222
overfitting, 179, 188, 198, 201
overlay names, 340
overload, cognitive, xv
overloading, 68
Owen, Melody, xviii

p-values, 139, 143, 161, 167, 181, 472
    perils, 165
    reporting, 166
Pacific time zone, 288
packages
    conflicts, 437
    guide, 435
    help, 438
    R, 434
    remove from workspace, 437
    used in book, 418
    versions, 457
packrat, 438, 457
Page, Larry, 390
PageRank, 378, 390, 391, 394, 397
pairwise relationships, 48
palettes, color, 19, 341
Pandas, 413
Pandoc, 454
panel background, 251
panel grid major, 251
papers in data science, 360
Parade magazine, 237
paradigm of science, 453
paradigm shift, 409
paradox, Simpson's, 165
parallel computation and high-performance computing
    task view, 414
parallel computing, 404, 406
parallel planes, 469
parallel processing, 403
parallel slope, 468
parameter
    conditional regression, 479
parameters
    conditional regression, 468
    regression, 463
    tuning, 180
parentheses, 368
parents, 48
parking lot, 331
parse number, 127
PARTITION defintion, *see* SQL
partitioning, 176, 188, 308
    recursive, 174, 198
party affiliation, 10, 212
passwords, 487
pathological variable names, 368
paths, 377
paths, garden of forking, 454
patterns, 149, 150
    counting, 356
patterns in data, 9
payroll, 227
PCA, 213, 218
pdf
    creating, 455, 457
    format, 454, 456
Pearson correlation, 189
peer review, 454
perceived threats, 132
percentages, 12
percentile, 151, 157
perception, graphical, 18
performance, 404
    query, 301
perils of extrapolation, 24
perils of p-values, 165
permission, read, 313
permutation testing, 5
Perrett, Jamis, 30
Perry, William (Refigerator), 28
persistent truth, 453
personal hygiene, 228
petabytes, 261, 401, 402
Peters, Arno, 324

phase transition, 379
PHI, 140
PHP, 243
phylogenetics, especially comparative methods task view, 436
physical experiment, 453
physician, 131, 317
π, 103
pie charts, 21, 41
Pig, 409
Pioneer Valley Planning Commission, 464
pipe operator, 4, 66, 73, 91, 444
pipeline, 262
  etl, 417
plaintiff, 136
planes, parallel, 469
plays, 356, 412
Plot.ly, 244
plotting text, 34
plurality, 182
Poisson distribution, 380
Poisson random variable, 232
Polar coordinates, 18
political action committees, 9
political parties, 215
political science, 142, 352
polygons, 319, 349
polynomial time solution, 173
Poole, Jane, 97
population, 33, 149, 151, 334
population parameters, 463
popups, 332
position, 21, 37
PostgreSQL, 267, 411, 485
potential conflict of interest, 137
POTUS, 378
power law, 371, 380
PowerPoint, 30
precipitation, 469
precursors to data science, 401
predicted values, 463, 474
prediction error, 189
predictions, 171, 192
predictive analytics, 8, 171, 173, 188, 201, 205
predictive model, 172
predictor variable, 463
predictors, categorical, 467
preferential attachment, 380
preprints, 360
presentations, 27, 30
  in RStudio, 457
President Obama, 10
presidential elections, 9, 333
presidents, 7, 10, 63
pressure
  air, 411
  blood, 99
prices, home, 141
primary key, 303
PRIMARY KEY definition, *see* SQL
Principal Component Analysis (PCA), 218
principal component analysis (PCA), 213
principles, 131
privacy, 131, 138
  policies, 140
private key, 369
private repositories, 457
privileged communications, 140
privileges, SQL, 488
probability, 46, 150, 173, 190, 476
  conditional, 183
probability distributions task view, 436
probe, genetic, 222
problem-solving, xv, 441
processors, 405
product-moment correlation, 189
productivity, xv
productivity, economic, 33
professional ethics, 131, 136, 143
programming, 3
  defensive, 450
programming interfaces, 117
programming, literate, 454
PROJ.4 string, 326
Project Gutenberg, 355, 374
Project TIER, 457
projection, 324, 328
  Albers, 324
  Gall–Peters, 324
  Lambert, 324
  Mercator, 324, 350
projections, 352
projections, maps, 352
projects, 421, 457
proportional symbol map, 344
proportions, 41
prostate cancer, 222
protected health information (PHI), 140
protocol, 139
  clinical trial, 144
  CONSORT, 144, 167

protocols, 166
provenance of data, 116
Pruim, Randall, xviii
pruning, 179
pseudo-random number seeds, 235
psychometric models and methods task view, 436
puberty, 471
public data set, 135
public health, 91, 328
public policy, 138
public repositories, 457
published studies, 454
Puerto Rico, 334
pump, 318, 328
punch cards, 3
Pythagorean Winning Percentage, 110
Python, 127, 404, 406, 408, 413

qq plot, 474
qualitative palette, 20
quantifying patterns, 149
quantile–quantile plot, 474
quantiles, 157
  t-distribution, 443
quantitative
  outcome, 475
  variable, 96
Quayola, 28
query language (SQL), 261
query optimization, 263
query performance, 301
query, search, 390
queues, 231
quitting R, 423
quotation marks, 368

R, xv, 404, 413, 421
  accessing databases, 490
  accessing variables, 447
  command history, 422
  data structures, 426
  Development Core Team, 421
  documentation, 438
  environment, 446
  exiting, 422
  FAQ, 425
  Foundation for Statistical Computing, 421
  help system, 423
  history, 421
  index, xvi
  installation, 422
  introduction, 421
  libraries, 434
  Linux installation, 422
  Markdown, 142, 244, 454, 457
  Markdown cheat sheets, 457
  objects, 426
  packages, 434, 438, 443
  programming, 425
  questions, 425
  resources for new users, 423
  sample session, 422
  starting, 422
  style guide, 104, 431
  task views, 435
  warranty, 423
  Windows installation, 422
R^2, 189, 465
race, running, 97
radial coordinate system, 22
rail trail, 464
rain, 469
RAM, 265, 404
Ramirez, Manny, 82
random chance, 167
random classifier, 191
random forest, 181, 200
random graph, 233, 378, 379
  Erdős–Rényi, 379
random noise, 149, 227
random number, 266
  normal, 227
  seed, 235, 445
  uniform, 224
random sample, 149, 175, 223
random selection, 150
randomized trials, 162, 166, 167
rank, 212
rank choice voting, 97
rank correlation, 189
rate limit, 370, 372
rate, false discovery, 223
ratio
  data to ink, 41
  of name frequency, 103
RBI, 82
reactive objects, 248
read permission, 313
readable files, 454
reading data, 116

Reagan National Airport, 285, 308
real estate database, 141
real-world case studies, xv
receiver operating characteristic curves, 189
record for running, 118
records, world, 21
rectangular array, 93
recursion, 174
recursive partitioning, 174, 198
Red Sox, 83
reduction, dimension, 211
reference category, 467
reference system, coordinate, 325
referencing variables, 190
referential integrity, 82, 304
regional airline, 292
registries, clinical trial, 144
regression, 44, 161, 163, 171, 172
    assumptions, 473
    coefficients, 474
    inference, 472
    interaction, 470
    logistic, 162, 173, 193, 475, 476
    model, xvi, 465
    multiple, 411, 463, 468
    parameters, 463
    regularization, 201
    residuals, 463
    ridge, 192, 201
    simple linear, 463
    tree, 173
regular expressions, 193, 253, 254, 355, 356, 374
regularization, 192, 201
rehearse, 27
reidentification, 138, 140, 141
Reinhard, Carmen, 135, 138
reject null, 166
relational calculus, tuple, 267
relational database, 98, 117, 261, 266, 303
relationships between variables, 12
relationships, pairwise, 48
relative humidity, 411
relevance, 390
reliability, 139, 154, 157
remove
    data frame from workspace, 431
    objects, 427
    package from workspace, 437
renaming variables, 68, 190
repeat statement, 423
repetitions, 359
REPLACE statement, *see* SQL
replication, 453
    crisis, 453
reporting p-values, 166
repositories, *see* GitHub
reproducibility, 234
reproducible analysis, xvi, 92, 142–144, 421, 454, 457
    packages, 438
    spreadsheets, 135, 138
    task view, 455
    workflow, 116, 457
reproducible research, 142, 167, 453, 457
reproducible research task view, 436
Republican, 10, 65, 335
resampling, 5, 115, 155, 223
research
    reproducible, 457
    responsible conduct, 143
reserved commands, 423
reset values, 308
reshape, 100, 101
residual diagnostics, 474
residuals, 474
response variable, 172, 205, 463
responsible conduct of research, 143
restaurant inspections, 7, 228
result, exact, 226
retaliation, 140
retraction, 142–144
retrieve old versions, 457
Retrosheet, 402
return value, 443
retweet counts, 371
Reuters, 132
revision control, *see* GitHub
rho, 189
Rickey, Branch, 6
ridge regression, 192, 201
right join, 281
Ripken, Cal, 87
Riseman Sisters, xviii
rising ocean levels, 134
Rmd file, 456
RMSE, 189
road maps, 208
Robinson, David, 374
Robinson, Jackie, 6
robust statistical methods task view, 436
ROC curves, 189

Rogers Commission, 23
Rogoff, Kenneth, 135, 138
Romney, Mitt, 10, 14, 333
root mean squared error (RMSE), 189
root privilege, 486
rOpenSci, 127
Rosling, Hans, 91
Roth, Allan, 6
rounding, 190, 472
routes, commuting, 331
Rouzer, David, 335
row, 97
rownames, add, 181
RSeek, 425
RStudio, xv, xviii, 243, 421, 455
  cheat sheets, 58, 88, 258, 457
  curated guide to learning R, 425
  developers, xviii
  Environment tab, 446
  installation, 422
  Packrat projects, 438
  presentations, 457
  reproducible analysis, 457
Rubin, Ben, 28
Ruby, 243
rules for naming, 103
rules of baseball, 6
run ratio, 110
running average, 444
running race, 97
  world record, 118
runs batted in (RBI), 82
runs scored and allowed, 73, 110
Ruth, Babe, 6
Ryan, Paul, 135
Rényi, Alfréd, 378

S3 generic functions, 417
S4 objects, 190, 321
sabermetrics, 6, 110
safety of patrons, 228
salary, teacher, 43, 162
sales, houses, 120, 141
Salt Lake City, UT, 346
sample, 149, 151, 223
  size, 150, 153, 234
  statistics, 152
  with replacement, 155
sampling distribution, 152, 153
San Francisco, 141, 150, 160
Saratoga, NY, 120
SAS files, 116
SAT scores, 20, 41, 162
  percentage taking, 43
save
  files, 116
  R history, 422
Scala, 413
scalability, 314
scalar, 105
scale, 15, 18, 37
  fill, 477
  vs guide, 18
scaled percentages, 12
scatterplot, 13, 43, 162
  jitter points, 476
  smoother, 470
scheduled departure time, 159
schema, 303, 310, 495
scheme, color, 340
Scholastic Aptitude Test, *see* SAT scores
Schuyler Sisters, 59
science
  network, 377
  of learning from data, xv
scientific method, 453
scikit-learn, 413
scoping, 447
score, 392
Scotland
  King, 359
  Parliament, 212
  political parties, 215
scraping data, 116, 369
scraping text, 367
scriptable computing, 454
scripts, 489
search box, 244
search engines, 365
search, Web, 390
second, 85
security, root privilege, 486
seeds, random number, 235
seeds, tournament, 391
segments, 346
SELECT statement, *see* SQL
selection bias, 464
self-intersections, 377
Senate, 7
sensor
  laser, 464
  network, 3

sequential palette, 20
server
    database, 485
    installation, 301
    MySQL, 262
    version of RStudio, 421
server file, 248
session information, 435
SET clause, *see* SQL
set inclusion operator, 91
setting up database server, 301, 485
sewer type, 120
SFO, 150
Shakespeare, 355, 356, 412
    Machine, 28
Shalizi, Cosma, 167
shape, 153, 324
shapefiles, 317, 319, 336, 352
sharing results, 143
Shea Stadium, 331
shell, 486
Shiny, 247, 332, 421
    cheat sheets, 258
    deploying, 250
    gallery, 258
shortcomings of PowerPoint, 30
SHOW INDEXES statement, *see* SQL
show indices, 305
show keys, 303
SHOW KEYS statement, *see* SQL
show tables, 494
shp files, 320
shrinkage, 201
shuffling, 223
shx files, 320
significant result, 144
Silge, Julia, 374
Silver, Nate, 290
similarity, 205
Simpson's paradox, 165
simulation, 150, 221, 231, 236
    null hypothesis, 165
    principles, 233
single, 85
singular value decomposition (SVD), 213
six degrees, 382
size, 34
size of an effect, 167
skills
    analytical, xv
    communication, xv
Skywest, 294
Slate.com, 141
slides in RStudio, 457
slope, 15, 463, 464, 468, 471
    different, 470
    parallel, 468
slots, S4 objects, 190
small multiples, 19
small world, 382
Smith College, xvii, 7, 330, 373, 411, 418
Smith, Will, 382
smoother, 43, 470
smoother, kernel, 41
snake case, 68
snow, 469
Snow White and the Huntsman, 389
Snow, John, 317
social hops, 382
social science, 457
Social Security Administration, 49, 93
socially conservative, 216
society, 143
Socrata, 127
software tools, 453
song
    titles, 367
    writers, 368
sorting, 69, 305
sorting arrows, 244
Sosa, Gabriel, xviii
source code control, *see* GitHub
source file, 454
space shuttle, 23
Spark, 409, 414
sparse matrix, 365
spatial data, 5, 352
spatial projection, 324
spatial statistics, 317
speaking line, 356
Spearman correlation, 189
special cases, 441
special characters in variable names, 127
specifying problems, 441
speculation to data, 221
spheroid, 324
spider trap, 397
spidering, 141
split string, 355
splitting, 176
spread, 152
spread data, 98

spreading data, 100
spreadsheets, 6, 265, 402, 454
  Excel, 91, 117, 139, 206
  Google, 91, 117
  perils of, 453
  reproducibility, 135, 138
SPSS files, 116
SQL, xvi, 63, 261, 262, 345, 404, 411, 418, 485
  administration, 301
  ALTER TABLE statement, 303
  AND operator, 274
  and R, 298
  baseball, 402
  BETWEEN operator, 273
  CHANGE clause, 303
  CONCAT function, 264, 273
  CREATE DATABASE statement, 302, 313
  CREATE TABLE statement, 302, 311
  creating databases, 301
  DESCRIBE statement, 268, 302
  dialects, 267
  DROP TABLE statement, 312
  efficient, 280
  errors, 313
  EXPLAIN statement, 306, 307, 491
  FOREIGN KEY definition, 303
  FROM clause, 263, 269, 307
  GROUP BY clause, 263, 269, 275
  HAVING clause, 263, 269, 278
  IF EXISTS clause, 312
  in a nutshell, 314
  IN operator, 274
  indexing, 273
  INSERT statement, 309
    IGNORE clause, 309
  install, 301
  JOIN clause, 263, 269, 281, 283
    CROSS JOIN clause, 281, 382
    LEFT JOIN clause, 281, 285, 307, 382
    OUTER JOIN clause, 281
    RIGHT JOIN clause, 281
  KEY definition, 303
  keys, 303
  LIMIT clause, 269, 280
  LOAD DATA
    LOCAL INFILE clause, 312
  LOAD DATA statement, 309, 312
  Microsoft, 267
  MongoDB, 267
  normal forms, 285
  OR operator, 274
  Oracle, 267
  ORDER BY clause, 263, 269, 277
  PARTITION definition, 308
  PostgreSQL, 267
  PRIMARY KEY definition, 303, 311
  privileges, 488
  REPLACE statement, 309
  scalability, 314
  SELECT statement, 263, 269, 270, 307, 382, 491
  SET clause, 308
  SHOW
    CREATE TABLE statement, 303
    DATABASES statement, 301
    INDEXES statement, 305
    KEYS statement, 303
    TABLES statement, 268
    WARNINGS statement, 312
  SQLite, 267
  str_to_date function, 273
  subqueries, 287
  table alias, 283
  TERMINATED BY clause, 312
  translation, 263
  UNION statement, 286
  UNIQUE KEY definition, 303
  universe, 266
  UPDATE statement, 308, 491
  USE statement, 302
  VALUES clause, 309
  WHERE clause, 263, 269, 272, 307
SQLite, 267, 372
squared residuals, 463, 465
SSE (sum of squared residuals), 465
SSM (sum of model squared deviations), 465
St. Joseph's, 397
St. Clair, Katie, xviii
stack overflow, 425
stakeholders, 137, 138, 143
stand your ground law, 132
standard deviation, 152
  test, 138
standard error, 153, 157
standard evaluation, 448
standards, ethical, 131, 143
standards, professional, 136
Stanford University, 243, 390

Starr, Ringo, *see* Beatles
Stata, 457
  files, 116
static graphics, 28
stationary distribution, 391
statistical computing, 236
statistical genetics task view, 436
statistical learning, 171, 205
statistical methodology, 149
statistical methods, 3
statistical modeling, 159
statistical programming, 451
statistical significance, 161, 165
statistically significant, 144
statistics, xv, 5, 149, 152
  definition, 152
  how to lie with, 131
  theoretical, 167
statistics for the social sciences task view, 436
step into function call, 450
steroid era, 86
Stewart number, 389
Stewart, Kristen, 382, 387, 389
Stonebraker, Michael, 267
storage mode, 431
storage, data, 141
storms, 134
straight line, 474
stratification, 165
streamgraphs, 246
strength of association, 149
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, 167
string, 125
  character, 355
  data, 355
  factors, 125
  split, 355
  to date, 273
STROBE statement, 167
Strogatz, Steven, 380
structured query language (SQL), 261
studies, published, 454
Sturgeon, Nicola, 217
style guide, *see* Google
subgroups
  iteration, 110
subject index, xvi
subqueries, 287
subsets, 173
Subversion, 457
summarizing, 70
  rows, 275
super-pac, 9
superuser, 488
supervised learning, 171–173, 201, 205
Supreme Court, 9, 138
survival analysis task view, 436
Sutton, Betty, 13
SVD, 213
Sweave, *see* knitr
swim records, 21
swimmer, 28
swing state, 333
swirl interactive courses, 425
symmetry, 216
syntax highlighting, 421
systolic blood pressure, 98

t-distribution, 472
  quantile, 443
tables
  alias, 283
  data, 244
  ephemeral, 288
  HTML, 118
  life, 49
  multiple, 98
  tabulate, 151
  versus graphs, 26
tabular content, 118
tabulate, 3, 151
talent, artistic, 28
tall data, 92, 98, 100
Tampa Bay Rays, 83
target time, 294
tartan, 212
task view, 127, 435
  analysis of ecological and environmental data, 436
  analysis of pharmacokinetic data, 436
  analysis of spatial data, 436
  Bayesian inference, 436
  chemometrics and computational physics, 436
  clinical trial design, monitoring, and analysis, 436
  cluster analysis and finite mixture models, 436

design of experiments (doe) and analysis of experimental data, 436
differential equations, 436
econometrics, 436
empirical finance, 436
extreme value analysis, 436
graphic displays and dynamic graphics and graphic devices and visualization, 436
graphical models in R, 436
handling and analyzing spatio-temporal data, 436
high-performance and parallel computing with R, 414, 436
machine learning and statistical learning, 218, 436
medical image analysis, 436
meta-analysis, 436
multivariate statistics, 436
natural language processing, 436
numerical mathematics, 436
official statistics and survey methodology, 436
optimization and mathematical programming, 436
parallel computation and high-performance computing, 414
phylogenetics, especially comparative methods, 436
probability distributions, 225, 436
psychometric models and methods, 436
reproducible research, 436, 455
robust statistical methods, 436
statistical genetics, 436
statistics for the social sciences, 436
survival analysis, 436
time series analysis, 436
Web technologies and services, 436
tau, 189
taxonomy of graphics, 9, 14
teacher salary, 43, 162
teaching
data science, 414
integrity in empirical research (Project TIER), 457
reproducible research, 457
statistics, 88, 413
team sport, xvi, 457
technical ability, 28
technical review, 142
technologies, data, xv
tellers, bank, 231
temperature, 24, 172, 411, 464
templates, markdown, 455
Temple University, 396
temporary file, 372
terabytes, 261, 401
term matrices, 365
terms, 390
of presidents, 66
of use, 138, 141
test
characteristics, 190
hypothesis, 165
test data, 175, 188
tests, 4
hypothesis, 472
reporting, 166
standard deviation, 138
unit, 451
Texas, 351
text
as data, 355
corpus, 364, 408
data, 5
ingesting, 367
mining, 28, 141, 355, 356, 374
theft, 141
themes, 250
Theorem, Bayes, 183
theoretical statistics, 167
Theron, Charlize, 389
thinking
computational, xv
with data, xv
Thorpe, Jer, 28
three V's, 401
threshold, 167, 228
threshold function, 379
tick marks, 38
tidy data, 91, 93, 95, 127
format, 116
multiple tables, 98
rules, 94
rules for, 93
tidyverse, 88, 332, 403
tilde, 172
tiles, 327
time of departure, 159
time scale, 18
time series analysis task view, 436
time variables, 123, 276

time zone, 288
times and dates, 272
timing commands, 403
title, 52
Tolstoy, Leo, 116
tools, software, 453
top billing, 383
topics on twitter, 373
Toronto, 287
tortured shapes, 341
tournament, 391
Toyota, 206, 210
track changes, 457
tradeoff
- database vs. disk, 265
- variance and bias, 192, 471

Tragedy of Macbeth, 355
trail crossings, 464
training data, 174, 175, 188
transactions, 231
transforming data, 93, 94
transition, 391
transition, phase, 379
translating
- integer codes, 121
- to SQL, 263

translation, 404
transpose, 100, 101
transposing data, 98, 100
trap, spider, 397
travel policy, 150, 157
tree
- aggregated, 181
- branches, 179, 205
- conditional inference, 201
- decision, 173
- evolutionary, 172, 205
- pruning, 179
- regression, 173

trend line, *see* smoother
trending topics, 373
triadic closure, 380
trials
- randomized, 162, 167

triple, 85
trucks, 206
true coefficients, 472
true positive, 191
truncate output, 280
truth, 195
truth, persistent, 453
truthful falsehoods, 131
Tufte, Edward, 23, 26, 30, 253
Tukey, John, 30
tuning parameters, 180
tuple relational calculus, 267
turnout, 94
Twilight, 48
Twitter, 7, 369, 371, 377, 378
type coercion, 428

UGESP, 134, 138
UMass, 393, 394, 396
undergraduate guidelines, 451
underscores, 104
undirected edges, 377
uniform distribution, 224, 397
Uniform Guidelines on Employee Selection Procedures (UGESP), 134, 138
UNION statement, *see* SQL
Unionist and Conservative Party, 215
unique, 155
unique cases, 97
unique key, 303
UNIQUE KEY definition, *see* SQL
unit tests, 451
United Airlines, 158, 294
United Kingdom Ordnance Survey, 326
United Nations, 143
univariate analysis, 24
univariate displays, 39
University of Auckland, 421
University of California, Berkeley, 144
University of Michigan, 144
University of North Carolina, 142
University of Washington, 406
UNIX-like shell, 485
unmeasured outcome, 172
unnamed arguments, 443
unsupervised learning, 171, 205, 218
update database, 308
UPDATE statement, *see* SQL
URL, 118
user accounts, 486
user interface, 248

Vaidyanathan, Ramnath, 243
validation, cross, 183, 188
valuations, houses, 141
VALUES clause, *see* SQL
varchar fields, 303

variability, 152
variable, 93, 96
  categorical, 96
  dependent, 463
  factor, 467
  independent, 463
  indicator, 467
  quantitative, 96
  renaming, 190
variable names
  pathological, 368
  special characters, 127
variance
  bias tradeoff, 192, 471
  residuals, 474
variety, 401
vector
  character, 355, 447
  extract elements, 429
  indexing, 427
  numeric, 442
  operations, 104
  recycling, 427
velocity, 401
Venn diagrams, 46
verb, 448
version control, 454, 457
  packages, 457
version number, 435, 438
vertices, 48, 233, 377, 384, 391
vetting, 142
victims, 134
victory, margin of, 392
view task, *see* task views
violations, health, 228
Vioxx, 135, 139
Virgin America, 158
Virgin Islands, 334
Virginia Tech, 397
visual cues, 15, 33, 56
Visual Display of Quantitative Information, 30
visualization, 3, 9, 149, 317
  graphs, 385
  interactive, 243, 258
  multivariate, 43
volume, 401
vos Savant, Marilyn, 237
voting, 97, 390

Wagaman, Amy, xviii
Wang, Susan, xviii
warranty for R, 423
Warren, Elizabeth, 14
Washington National Airport, 285
Washington, D.C., 97, 332
Watergate, 65
Watts, Duncan, 380
weather data, 7, 418
weave, *see* markdown
Web applications, 247, 421
Web comic, 253
Web Crawler, 390
Web search, 390
Web technologies, 127
Web technologies and services task view, 436
website for book, xvii
weekday, 470
weighted edges, 377
Weka machine learning library, 201
Weld, William, 140
WHERE clause, *see* SQL
while statement, 423
White House, 332
Wickham, Hadley, xviii, 30, 33, 58, 63, 88, 93, 104, 116, 413, 438
wide data, 98, 100
widgets, 248
Wikipedia, 118, 367
Wilkinson, Leland, 30, 58
Williams, Robin, 101
Windows, 485
  installation of R, 422
  R FAQ, 422
winning percentage, 73, 110
witness, expert, 140
women's names, 56
Word
  format, 454–456
  Microsoft, 454
word frequency, 406
workflow, xv, 401
  reproducible, 116, 454, 457
workspace, 437
  browser, 421
  conflicts, 431, 437
world records, 21, 118
World Wide Web, 391
wrangling, 3, 88
  cheat sheets, 88
  data, xv, 63, 94

tidy data, 91, 95
times and dates, 272
wrangling data, 447

xkcd, 253
XML files, 117, 351

Yahoo, 390, 402
Yankees, 112
Yau, Nathan, 14, 53, 254
year, month, day format, 123

Zillow, 141
ZIP file, 206
zodiac sign, 135