Index

Symbols

* asterisk MDX cross join, 495
{ } braces for sets in MDX, 491
: range operator MDX, 508
ε, 381

A

absolute value, 222
actions, 146
addition rule, 166
Adventure Works, 142
agglomerative clustering, 231
aggregation, 58, 142, 143
Agrawal, R., 276
Akaike information criterion (AIC), 395
Alberta, 217
Algorithm Parameters, 409
Amazon, 270
Analysis database, 490
analytics, 3
AND, 54
AnnTaylor Stores Corp., 13
anomaly, 205
antecedent, 273
application programming interface (API), 469
a priori algorithm, 16, 283
a priori tool, 293
a priori value, 175
ArcGIS, 468
ARIMA, 378, 425
  differencing, 390
  lag coefficients, 390
arithmetic operators, 57
ARMA(p, 390
ARMAX, 421
AR(p), 381
arrangements, 164
ARTxp, 406, 419, 425
  forecast, 415
  model, 413
association, 271
association analysis
  continuous data, 288
  data, 290
  quantity differences, 289
  traditional tools, 292
association rules, 273, 294
attribute, 220
  categorical, 223
  ordinal, 223
Attribute Discrimination, 348
attribute evaluation, 338
  neural network, 362
attribute relationship, 126, 130
Aunt Bessie's, 376
autocorrelation function (ACF), 391
auto regression, 377, 381
auto regressive integrated moving average (ARIMA), 389
auto regressive moving average (ARIMA)
  traditional tools, 397
auto-regressive tree with cross prediction (ARTxp), 406
average, 385
axes
  cube, 489

B

Barabba, Vincent P, 37
bar-code scanners, 9
Bayesian, 162
Bayes Theorem, 172, 173, 341
  a priori value, 175
Bayes, Thomas, 172
best and worst, 65
Best Buy, 11
BETWEEN, 51, 56
big data, 3, 16
Bing, 468, 472
BinomDist, 180
binomial distribution, 177
bins, 437
BLAST, 437, 458
Boolean algebra, 53
bootstrap, 196
bootstrapping, 25, 358
Borgelt, Christian, 293
BottomCount, 504
Box and Jenkins, 419
Box-Jenkins, 378, 389
Brat, Ilan, 37
brute force, 16
Bryon, Ellen, 37
BULK INSERT, 447
Bureau of Economic Activity (BEA), 393
business intelligence, 3
Business Intelligence Development Studio, 109

C

C#, 61
calculated measures
  MDX, 495
calculation, 134
CASE, 153, 317
cases, 30
  Bakery, 32
  basketball, 32
  Cars, 32
  Corner Med, 31
  diner, 31
  NBA, 32
  Rolling Thunder Bicycle Company, 30
Cast, 312
categorical attribute, 223
categories, 437
causality, 308
causation, 27, 187, 322
Census Bureau, 459
  demographic data, 434
central limit theorem, 195
CGAT, 435
chaotic, 381
ChiDist, 194
children, 484
Chi-square distribution, 194, 204, 395
Chi-square hypothesis tests, 203
classification, 309
  sequences, 436, 439
Classification Matrix, 340, 356
cluster

hierarchical, 229
sequences model, 438
cluster analysis, 217
Cluster Characteristics, 245, 453
CLUSTER_COUNT, 249, 451
Cluster Diagram, 244
Cluster Discrimination, 453, 454
clustering, 217, 218, 345
categorical data, 258
data, 236, 239
discrete data, 237
Microsoft, 241
missing data, 238
sequences, 436
Weka, 260
CLUSTERING_METHOD, 249
ClusterProbability, 247
Cluster Profile, 244, 453
CoalesceEmpty, 503
columns, 43, 489
combination, 165
combinatorial search, 224
comma-separated values (CSV)
SQL Server Management Studio, 399
comma-separated-values (CSV), 22, 30, 237, 446
comparison
time series models, 425
concatenation, 321
conditional color, 135
conditional probability, 171, 341
continuous, 182
confidence, 277
confidence interval, 198
consequent, 282
contingency table, 170, 341
continuous data, 163, 181
CONVERT, 51
correlation, 308, 463
pair-wise, 254
correlation coefficient, 186, 223
correlogram, 400
residual, 402
counting
order does not matter, 165
order matters, 164
counting and combinations, 163
covariance, 344
CREATE TABLE, 446
CREATE VIEW, 69
critical value, 202
cross correlation, 380, 392
linear regression, 417
cross join, 49, 494
cross-support, 287
cross validation, 27, 325
cube, 482
actions, 146
aggregations, 142, 143
calculation, 134
partitions, 142
perspective, 134
structure, 120
cube browser, 81, 96, 102
SSAS, 119
cumulative distribution, 179
cumulative distribution function (cdf), 180
currencies, 142
curse of dimensionality, 282
customer behavior, 440
customers, 218
cyclical, 379

D

Dallas Cowboys, 481
dangers, 24
bad data, 25
estimation instability, 28
human error, 24
insufficient data, 25
model instability, 29
over fitting, 26
Dartmouth Atlas of Health Care, 463
data, 3, 16
associations, 271
smoothing, 384
database, 17, 42
Corner Med, 258
design, 82
relational, 43
structure, 46
Database
Bakery, 272
cars, 238
database design
first normal form, 86
second normal form, 86
third normal form, 87
database management system (DBMS), 3, 42
data definition, 72
data manipulation, 72
Data Mapping Wizard, 465
data mining, 3
goal, 310
data reduction, 233
data source, 240
data source view, 241
data type, 45
data warehouse, 19, 43, 99
models, 101
decision tree, 308, 309, 313, 320, 323, 349, 406, 419
data, 352
degrees of freedom, 204
DELETE, 74
demographic data, 466
dendrogram, 232
dependent, 311
dependent variable, 27, 423
DESC, 52
detail section, 79
diagnostic tool, 387
differencing, 390
digital dashboard, 96, 481
dimension, 102, 120, 483
evaluation, 309
hierarchy, 103, 107, 483
nominal, 223
problem of size, 282
reduction, 233
structure, 132
time, 124
dimension evaluation
data, 311
discrete, 163, 444
discrete data, 176
discretize, 437
discretized data, 313
discretizing, 163

distance, 222
DISTINCT, 58, 444
divisive clustering, 230
DNA, 435, 442
drill down, 81, 95, 108
drill through, 147, 354
drop-down list, 81
Duhigg, Charles, 37
dummy variable, 381, 423
Dvorak, Phred, 37

E

e-commerce, 9
econometricians, 316
edit distance, 436
eigenvalues, 235
elasticity, 327
employees, 12
endogenous, 311
enterprise resource planning (ERP), 19, 42, 376
entity-relationship diagram (ERD), 68
entropy, 205
error term, 381
Esri, 468
Euclidean measure, 222
evaluate dimensions, 309
Excel, 119, 144, 180
  linear regression, 318
EXCEPT, 501
EXECUTE, 446
exogenous, 311
expectation maximization (EM), 228
expected value, 183
experiment, 176
exponential, 368
exponential growth, 388
extensible markup language (XML), 46
extraction transformation and loading (ETL), 19
extraction, transformation, and loading (ETL), 100

F

fact, 102
Federal Aviation Administration (FAA), 41
federal government
  data sources, 473
finance, 4, 27, 220
first normal form, 86
forecast, 377, 378, 404, 415
forecasting, 3
FormatString, 121
four nines, 168
Friedman, Jerome, 38
FROM
  MDX, 489
function, 60
  length, 61
  Max, 65

G

gap
  seqences, 436
gap statistic, 227
Gaussian, 189
General Motors, 6
general multiplication rule, 171
Generic Content Tree Viewer, 362
geographic analysis, 435
geographic correlation, 460
geographic information system (GIS), 434, 459
global positioning system (GPS), 459
Goodman, Peter S, 38
goodness of fit, 203, 204
Google, 12, 459, 468, 471
Google Analytics, 448
Google map, 147
Göransson, H., 38
gretl, 22, 400, 419
GROUP BY, 61, 63, 319, 335, 396, 408, 487
Gustafsson, M.G., 38

H

Hastie, Trevor, 38
HAVING, 63
HIDDEN_NODE_RATIO, 361
hierarchical
  dimension, 483
hierarchical clustering, 230
hierarchical clusters, 229
hierarchy, 103, 107
  location, 128
  product, 274
  time, 123
high-frequency trading, 5
Hopper, Max, 3
Houston Pawn Shops, 434
HTML, 46
Hudson, Simon, 217
human biases, 160
human capital, 13
hybrid (HOLAP), 102
hyper cube, 3, 116
hypergeometric distribution, 178
hypotheses, 3
hypothesis testing, 200

I

identity, 44
Ilan, 37
immediate if function
  MDX, 502
immediate if function (IIF), 135
independent, 167, 311
independent variable, 27
index, 18, 98
information, 3, 174, 343, 351
  Bayes Theorem, 174
information measure, 205
INNER JOIN, 67
INSERT, 73
instability
  estimation, 28
  model, 29
interestingness, 277
internationalization, 140
IP address, 438
Isaksson, A, 38
itemsets, 277

J

Javascript, 469, 471
Johnson, Avery, 38
join

LEFT JOIN, 71
many tables, 68
JOIN, 66
joining tables, 66
joint events, 167
joint probability, 167, 169, 170
continuous, 182
just-in-time, 12

K

key column, 312, 321
key peformance indicators (KPI)
status expression, 152
key performance indicators (KPI), 96, 149
trend expression, 152
key value, 44
K-means, 224, 250
knowledge, 4
Koudsi, Suzarne, 38

L

lag, 381, 386
lag limits, 390
layer
map data, 460
learning
machine, 14, 16
leaves, 484
LEFT JOIN, 71
levels, 484, 489
dimension, 273
Levenshtein distance, 436
lift, 277
LIKE, 56
Lilienfeld, Scott, 160
linear regression, 308, 309, 313, 381, 422, 426
cross correlation, 417
goals, 314
Linear Regression, 313
local geography, 461
location-based data, 435
location hierarchy, 128
Logical Primary Key, 259
logistic regression, 308, 313, 330
Log Parser 2.2, 445
Lohr, Steve, 38

M

machine learning, 3, 16
MapPoint, 464
Data Mapping Wizard, 465
margin totals, 170
market basket, 16, 270, 271
market basket data structure, 291
marketing, 6, 220
Markov chain, 439
materialized view, 100
maximum likelihood estimator (MLE), 351
MAXIMUM_SEQUENCE_STATES, 451
McClelland, James L., 38
MDX, 482
calculated measures, 495
MDX function, 501
Avg, 508
CoalesceEmpty, 503
EXCEPT, 501
IIF, 502
NextMember, 498
ParallelPeriod, 499
PrevMember, 498
TopCount, 504
TopPercent, 504
TopSum, 504
YTD, 506
MDX structure
FROM, 489
SELECT, 489
WHERE, 489
WITH MEMBER, 489
mean, 191, 196
sample, 196
mean absolute deviation (MAD), 326
measures, 102, 195, 239, 482
continuous, 163
discrete, 163
member
MDX, 489
metadata, 102
Microsoft Generic Content Tree Viewer, 456
Microsoft MapPoint, 460, 464
Miner3D, 240
minimum confidence, 284
minimum support, 284
Mining Accuracy Chart, 324, 337
Mining Legend, 244, 411
Mining Models, 249
Mining Model Viewer, 362
Mining Structure, 321, 353
missing data, 238, 313, 317
mixture model, 227
model, 4, 12, 29, 309
association analysis, 274
data warehouse, 101
time series, 379
model comparison, 365
Morgenson, Gretchen, 38
Morrison, Scott, 38
moving average, 384
MDX, 507
multicollinearity, 29, 233
multidimensional expression (MDX), 150
multidimensional OLAP (MOLAP), 101
multi-dimension expression (MDX), 482
multinomial, 333
multiple tables, 66
multiplication rule, 167
general, 171
multivariate normal distribution, 192
mutually exclusive, 166
MySQL, 22

N

Nabisco, 9
naïve Bayes, 308, 313, 340
named calculation, 112, 113
named query, 114, 320
National Center for Biotechnology Information, 458
National Football League (NFL), 481
National Institute for Health, 458
natural language, 43
Netflix, 9, 32, 269

competition, 10
neural network, 16, 308, 313, 358, 359
data, 361
goals, 359
Neural Network Model Viewer, 338
neuron, 359, 360
NextMember, 498
node, 321
nominal, 223
NON EMPTY
MDX, 492
nonlinear
complications, 368
nonlinear dangers, 389
nonlinear relationships, 16
nonlinear trend, 387
normal distribution, 189
normalization, 82
NormDist, 190
NOT, 54
NULL, 57, 313
null hypothesis, 201

O

O'Connell, Vanessa, 38
ODBC, 22
OLAP
snowflake design, 107
star design, 105
OLAP cube, 396, 482
one-to-many relationship, 83
online analytical processing (OLAP), 19, 96, 482
online transaction processing (OLTP), 18, 98
opacity, 469
openstreetmap.org, 473
OR, 55
ORDER BY, 52, 115
order of operations, 136
ordinal measure, 223
Oreo, 9
orthogonal, 233
over fitting, 26, 196

P

ParallelPeriod, 151, 499
parallel processing, 24
parameter, 183
auto regression, 383
parent, 484
values in MDX, 497
partial autocorrelation function (PACF), 391
partitions, 142
pattern matching, 436
peak load, 179
percentage, 137
MDX, 497
performance
OLAP cube, 142
periodicity, 412
permutation, 164
perspective, 134, 138
Pfizer, 7
PivotChart, 144
PivotTable, 119, 144, 494
Poisson distribution, 179
polynomial, 368
posterior distribution, 343
PostgreSQL, 22
prediction, 14, 246, 309, 366, 440, 457
decision tree, 355
logistic regression, 339
naïve Bayes, 348
neural network, 363
regression, 328
Prediction Function, 247
PREDICTION JOIN, 416
predictive software, 307
PredictNodeID, 416
PredictTimeSeries, 415
PredictVariance, 416
PrevMember, 498
primary key, 44, 66, 79, 87
principal components analysis (PCA), 233
prior distribution, 343
probability, 15, 161
addition rule, 166
conditional, 171
frequency, 166
joint, 170
multiplication rule, 167
relative frequency, 161
rules, 166
subjective, 162
probability density function (pdf), 181
probability distribution, 177
binomial, 177
chi-square, 194
hypergeometric, 178
normal or Gaussian, 189, 227
Poisson, 179
T (Student's T), 192
uniform, 188
probability distributions, 176
probability function, 177
probability mass function, 177, 181
problems
nonlinear, 389
Providian, 5

Q

Q-statistic, 395
quality assurance, 41
query, 47, 78, 101
basics, 49
editor, 50
four questions, 47
saved, 69
query-level calculation, 137
query system, 42

R

random, 161
random chance, 27
random error, 27, 386
random events, 161
random paths, 437
random sample, 195
random variable, 177
recommendation engine, 271

Reebok, 481
relational database, 43, 109
relational OLAP (ROLAP), 102
relationship, 111, 309
  attribute, 130
  data source view, 115
  one-to-many, 83
relationships, 15
relative frequency, 161
relative risk, 278
repeating section, 86
REPLACE_MODEL_CASES, 416
report
  break item, 79
  detail section, 79
  query, 78
  wizard, 80
Reporting Services, 75
reports, 75
report wizard, 76
residual, 402
responsibilities, 228
retail stores, 8
Ritchie, Brent, 217
ROLLBACK, 74
Rolling Thunder Bicycle Company, 44, 319, 345, 391, 463
roll up, 108
root, 484
root mean square error (RMSE), 326, 358, 395
row, 43, 489
row-by-row calculation, 58
R-squared, 323
R System, 22
rules, 270
Rumelhart, David E., 38

S

sample, 195
  mean, 196
  variance, 196
sample space, 176
sample variance, 196
Schwarz criterion, 396
seasonal ARIMA (SARIMA), 407
seasonal auto-regressive (SAR), 407
seasonal effect, 377
seasonality, 379
  evaluation, 405
seasonally adjusted, 394
seasonal moving average SMA, 407
second normal form, 86
SELECT
  DISTINCT, 444
  MDX, 489
SELECT, FROM, JOIN, WHERE, 52
sequence analysis
  data, 442
  missing data, 444
Sequence Cluster, 456
sequence clustering, 447
sequence mining, 436
sequences, 435
  classification, 439
  clustering, 452
set, 489
Shannon, Claude, 205, 351
Shannon's entropy, 351
Simpson, E.H., 285
Simpson's paradox, 285
skewed support, 286
slice, 492
slope coefficient, 315
Smith, Adam, 11
smooth data, 384
snowflake design, 107
software tools, 20
Solution Explorer, 321
spurious correlation, 287
SQL, 17, 42, 101, 487
  aggregation, 58
  AND, 54
  BETWEEN, 56
  CASE, 316, 332
  CASE statement, 113
  DELETE, 74
  DESC, 52
  DISTINCT, 58
  function, 60
  GROUP BY, 61, 63
  HAVING, 63
  INNER JOIN, 67
  INSERT, 73
  introduction, 52
  JOIN, 66
  LIKE, 56
  NOT, 54
  NULL, 57
  OR, 55
  ORDER BY, 52
  ROLLBACK, 74
  SELECT, 52
  subquery, 70
  UNION, 71
  UPDATE, 72
  view, 69
  WHERE, 58
SQL function
  CAST, 312
  Month, 397
  Year, 319, 397
SQL query, 299
SQL Server, 30, 43, 61
  Business Intelligence (BI), 75
  reports, 75
SQL Server Analysis Services (SSAS), 20, 21, 99
  OLAP cube, 108
SQL Server Business Intelligence (BI), 75
SQL Server Management Studio, 399, 490
squared-difference, 222
SSAS
  data sources, 109
  data source view, 111
  deployment and processing, 118
stability
  model, 327, 368
standard deviation, 185, 191
  mean, 199
star design, 105
state transitions, 456
stationary, 392
statistical mixture model, 227
statistical research, 3
statistical theory, 3
statistical tools, 22
  gretl, 22
  SAS, 22

SPSS, 22
Stata, 22
statistician, 25
statistics, 15, 161, 195
Status expression, 152
strings, 437
subjective, 162
subquery, 70
subtotal, 61, 119
sum of squared errors, 316
supply chain management, 11
support, 277, 279
minimum level, 283
surprise, 206
swindle, 5

T

table, 43, 48
join, 66
Tableau Software, 481
table join, 48
Taiwan Semiconductor Manufacturing (TSMC), 11
Taylor III, alex, 38
T distribution, 192
third normal form, 87
Tibshirani, Robert, 38
TIGER mapping system, 461
time dimension, 124
time key, 408
time series, 377
data, 396
missing data, 397
model, 379
TOP 5, 65
TopCount, 504
TopPercent, 504
TopSum, 504
transactions, 4
transition probability, 439
translation, 140
tree diagram, 171
trend, 379, 387, 399
nonlinear, 387
Trend Expression, 152
T statistic, 323
tuple, 489
Type I error, 201
Type II error, 201

U

Unicode, 45
uniform distribution, 188, 206
UNION, 71
union of events, 166
Union Pacific, 307
University of Waikato, 252
unsupervised learning, 20, 219, 270
UPDATE, 72
U.S. Geological Survey (USGS), 473

V

variance, 184, 185, 328
sample, 196
Venn diagram, 279
view, 69, 99
Vincent P., 37
visualization, 254, 481
Visual Studio, 21, 75, 109, 490
Visual Studio Designer, 508

W

Wallman, M., 38
Wal-Mart, 2, 32
Washington Mutual, 5
Web logs, 445
Web site traffic, 447
Web site usage, 436
weighted average, 385
Weka, 22, 238
WHEN, 153
WHERE, 58, 64
MDX, 489, 492
versus HAVING, 64
Winmetrics Corp., 95
wisdom, 4
within-cluster distance, 225
WITH MEMBER, 489
wizard
report, 80

X

XML, 45
XQuery, 46

Y

YTD, 506

Z

Zaltman, Gerald, 37
Zimmerman, Ann, 37
Zoran, 11
Z statistic, 203
Z value, 198
Zweig, Jason, 38