

References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 2nd edition, 2011.
2. Carlo Batini and Monica Scannapieca. *Data Quality: Concepts, methodologies and techniques*. Springer, 2006.
3. Richard Belew. *Finding Out About*. Cambridge University Press, 2008.
4. Michael Berthold, Christian Borgelt, Frank Höppner, and Frank Klawonn. *Guide To Intelligent Data Analysis*. Texts in Computer Science. Springer, 2010.
5. Tamraparni Dasu and Theodore Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley and Sons, 2003.
6. AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.
7. Xin Luna Dong and Divesh Srivastava. Big data integration. *Synthesis Lectures on Data Management*, 7(1):1–198, 2015.
8. Joseph M. Hellerstein. Quantitative data cleaning for large databases. Technical Report, UC Berkeley, 2008. <https://dsf.berkeley.edu/jmh/papers/cleaning-unece.pdf>
9. Jeroen Janssens. *Data Science at the Command Line*. O'Reilly, 2015.
10. Daniel Lemire and Anna Maclachlan. Slope one predictors for online rating-based collaborative filtering. *CoRR*, abs/cs/0702144, 2007.
11. Gordon Linoff. *Data Analysis Using SQL and Excel*. Wiley, 2008.
12. Christopher D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
13. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.
14. Kevin Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
15. David Skillcorn. *Knowledge Discovery for Counterterrorism and Law Enforcement*. Chapman and Hall/CRC Data, 2008.
16. Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to Data Mining*. Pearson, 2nd edition, 2019.
17. Robert Trueblood and John Lovett. *Data Mining and Statistical Analysis Using SQL*. Apress, 2001.
18. Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1), 2011.
19. Hadley Wickham. Tidy data. *Journal of Statistical Software*, 51(10), 2014.
20. Hadley Wickham and Garret Grolmund. *R for Data Science*. O'Reilly, 2017.

Index

A

Aggregation, 156
ALL, 227
ALTER TABLE, 41
Anomaly, 52
ANOVA, 125
ANY, 227
A priori, 195–197
Association Rule, 193–195
Association Rule, confidence, 194
Association Rule, support, 193

B

BETWEEN, 81
Big Data, 269
Bigrams, 209
Binarization, 156, 158
Binning, 115, 117–119

C

Cartesian product, 83
CASE, 93
Categorical Data, 19, 136
CHECK, 241
Chi square, 126, 127
Classification, 173
Cloud, 278
Cluster, 276, 277
Clustering, 186, 189, 235
COALESCE, 150
Contingency table, 126
Covariance, 122
CREATE DATABASE, 31

CREATE TABLE, 32, 35, 42
Crosstab, 163
CUBE, 236

D

Data Cleaning, 4, 132
Data Life Cycle, v, 1, 2
Data Loading, 69, 70, 72
Data mart, 275
Data Type, 32
Data warehouse, 271
Dates, 34, 140, 141, 143
DELETE, 72
Dice, 274
Discretization, 156, 157
Distance, 153, 186, 235
Distance, Cosine, 188
Distance, Euclidean, 187
Distance, Mahalanobis, 187
Distance, Manhattan, 187
DISTINCT, 82, 92
Drill-down, 275
Dummy Encoding, 163
Dumping data, 74
Duplicate data, 4, 82, 133, 152, 153, 155

E

Entropy, 119
ETL Process, 273
EXISTS, 226
Exploratory Data Analysis (EDA), 3, 77

F

Foreign Key, [53–55](#), [57](#)
FROM, [78](#)
Functions, Aggregate, [89](#), [91](#), [92](#)
Functions, Standard, [89](#)
Fuzzy matching, [153](#)

G

Graph Data, [14](#), [64](#)
GROUP BY, [156](#), [236](#)

H

Hadoop, [276](#)
Histogram, [114–116](#), [118](#)
Hive, [277](#)

I

IN, [80](#), [226](#)
Information Retrieval (IR), [202](#)
INSERT, [36](#)
Inverse Document Frequency (idf), [205](#)

J

JOIN, [83](#), [87](#), [221](#), [223](#)
JSON, [10](#), [61](#), [63](#)

K

Keys, [32](#), [38](#), [42](#)
Keyword search, [202](#), [206](#), [208](#)
K-means, [189](#)
kNN, [191](#)
Kurtosis, [107](#), [114](#)

L

LIKE, [80](#), [153](#)
LIMIT, [101](#)
Linear regression, [179](#)
Logistic regression, [184](#)

M

Many-to-many relationship, [54](#)
Matrix, [65](#)
Matrix, Adjacency, [65](#)
Mean, [107–110](#), [234](#)
Median, [107](#), [111](#), [233](#)
Metadata, [3](#), [24](#), [38](#)
Missing data, [4](#), [133](#), [144](#), [145](#), [148](#)

Mode, [107](#), [111](#), [234](#)
Multi-valued Attribute, [50](#)
Mutual information, [121](#)

N

Naive Bayes, [173](#)
Nearest Neighbors, [191](#)
Normalization, [5](#), [52](#), [134](#), [135](#)
NULLs, [50](#), [145–147](#)
Numerical data, [19](#), [21](#)

O

OFFSET, [102](#)
One-to-many relationship, [53](#)
One-to-one relationship, [53](#)
ORDER BY, [101](#)
Ordinal Data, [19](#), [21](#)
Outliers, [4](#), [133](#), [150](#), [151](#), [234](#), [235](#)

P

Pearson correlation, [122](#), [123](#)
Percentiles, [233](#)
Pivoting, [5](#)
Pointwise Mutual Information (PMI), [121](#)
Probability, Joint, [120](#), [121](#)
Provenance, [167](#)

R

Rank, [103](#), [202](#), [232](#)
Rank correlation, [123](#)
Renaming, [86](#), [87](#)
ROLLUP, [236](#), [274](#)

S

Sampling, [156](#)
Scaling, [134](#)
Schema, [7](#), [32](#)
SELECT, [77](#)
Semistructured data, [7](#), [9](#), [61](#)
Sentiment analysis, [211](#)
Skewness, [107](#), [114](#)
Slice, [274](#)
Snowflake schema, [272](#)
Standard Deviation, [107](#), [112](#)
Standardization, [5](#), [134](#)
Star schema, [272](#)
Strings, [32](#), [136–139](#)
Structured data, [7](#)
Subquery, [78](#), [79](#), [92](#), [225](#)
Supervised learning, [171](#), [172](#)

T

Table, [32](#)
Tabular data, [8](#)
Term Frequency (tf), [205](#)
Tidy data, [vi](#), [43](#), [44](#)
Tokenization, [203](#)
Top k, [101](#), [233](#)

U

Unstructured data, [7](#), [16](#), [67](#)
Unsupervised learning, [172](#)
UPDATE, [74](#)

V

Variance, [113](#)

W

WHERE, [78](#)
Windows, [229](#), [230](#), [232](#)
WITH, [103](#), [104](#)

X

XML, [10](#), [61](#), [63](#)