

Table of Contents

Preface.....	xiii
1. Exploratory Data Analysis.....	1
Elements of Structured Data	2
Further Reading	4
Rectangular Data	4
Data Frames and Indexes	6
Nonrectangular Data Structures	6
Further Reading	7
Estimates of Location	7
Mean	9
Median and Robust Estimates	10
Example: Location Estimates of Population and Murder Rates	12
Further Reading	13
Estimates of Variability	13
Standard Deviation and Related Estimates	14
Estimates Based on Percentiles	16
Example: Variability Estimates of State Population	18
Further Reading	19
Exploring the Data Distribution	19
Percentiles and Boxplots	20
Frequency Tables and Histograms	22
Density Plots and Estimates	24
Further Reading	26
Exploring Binary and Categorical Data	27
Mode	29
Expected Value	29
Probability	30

Further Reading	30
Correlation	30
Scatterplots	34
Further Reading	36
Exploring Two or More Variables	36
Hexagonal Binning and Contours (Plotting Numeric Versus Numeric Data)	36
Two Categorical Variables	39
Categorical and Numeric Data	41
Visualizing Multiple Variables	43
Further Reading	46
Summary	46
2. Data and Sampling Distributions.....	47
Random Sampling and Sample Bias	48
Bias	50
Random Selection	51
Size Versus Quality: When Does Size Matter?	52
Sample Mean Versus Population Mean	53
Further Reading	53
Selection Bias	54
Regression to the Mean	55
Further Reading	57
Sampling Distribution of a Statistic	57
Central Limit Theorem	60
Standard Error	60
Further Reading	61
The Bootstrap	61
Resampling Versus Bootstrapping	65
Further Reading	65
Confidence Intervals	65
Further Reading	68
Normal Distribution	69
Standard Normal and QQ-Plots	71
Long-Tailed Distributions	73
Further Reading	75
Student's t-Distribution	75
Further Reading	78
Binomial Distribution	78
Further Reading	80
Chi-Square Distribution	80
Further Reading	81
F-Distribution	82

Further Reading	82
Poisson and Related Distributions	82
Poisson Distributions	83
Exponential Distribution	84
Estimating the Failure Rate	84
Weibull Distribution	85
Further Reading	86
Summary	86
3. Statistical Experiments and Significance Testing.....	87
A/B Testing	88
Why Have a Control Group?	90
Why Just A/B? Why Not C, D,...?	91
Further Reading	92
Hypothesis Tests	93
The Null Hypothesis	94
Alternative Hypothesis	95
One-Way Versus Two-Way Hypothesis Tests	95
Further Reading	96
Resampling	96
Permutation Test	97
Example: Web Stickiness	98
Exhaustive and Bootstrap Permutation Tests	102
Permutation Tests: The Bottom Line for Data Science	102
Further Reading	103
Statistical Significance and p-Values	103
p-Value	106
Alpha	107
Type 1 and Type 2 Errors	109
Data Science and p-Values	109
Further Reading	110
t-Tests	110
Further Reading	112
Multiple Testing	112
Further Reading	116
Degrees of Freedom	116
Further Reading	118
ANOVA	118
F-Statistic	121
Two-Way ANOVA	123
Further Reading	124
Chi-Square Test	124

Further Reading	236
Summary	236
6. Statistical Machine Learning.....	237
K-Nearest Neighbors	238
A Small Example: Predicting Loan Default	239
Distance Metrics	241
One Hot Encoder	242
Standardization (Normalization, z-Scores)	243
Choosing K	246
KNN as a Feature Engine	247
Tree Models	249
A Simple Example	250
The Recursive Partitioning Algorithm	252
Measuring Homogeneity or Impurity	254
Stopping the Tree from Growing	256
Predicting a Continuous Value	257
How Trees Are Used	258
Further Reading	259
Bagging and the Random Forest	259
Bagging	260
Random Forest	261
Variable Importance	265
Hyperparameters	269
Boosting	270
The Boosting Algorithm	271
XGBoost	272
Regularization: Avoiding Overfitting	274
Hyperparameters and Cross-Validation	279
Summary	282
7. Unsupervised Learning.....	283
Principal Components Analysis	284
A Simple Example	285
Computing the Principal Components	288
Interpreting Principal Components	289
Correspondence Analysis	292
Further Reading	294
K-Means Clustering	294
A Simple Example	295
K-Means Algorithm	298
Interpreting the Clusters	299

Selecting the Number of Clusters	302
Hierarchical Clustering	304
A Simple Example	305
The Dendrogram	306
The Agglomerative Algorithm	308
Measures of Dissimilarity	309
Model-Based Clustering	311
Multivariate Normal Distribution	311
Mixtures of Normals	312
Selecting the Number of Clusters	315
Further Reading	318
Scaling and Categorical Variables	318
Scaling the Variables	319
Dominant Variables	321
Categorical Data and Gower's Distance	322
Problems with Clustering Mixed Data	325
Summary	326
Bibliography	327
Index	329

The goals of this book

- To present, in digestible, navigable, and easily referenced form, key concepts from machine learning that are relevant to data science.
- To explain when concepts are important and useful from a data science perspective, what to do next, and why.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic

Indicates new terms, URLs, email addresses, filenames, and file extensions.

Constant Width

Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, and keywords.