Preface. .................................................................................................................. xiii

1. Exploratory Data Analysis. ................................................................................. 1
   Elements of Structured Data 2
     Further Reading 4
   Rectangular Data 5
     Data Frames and Indexes 6
     Nonrectangular Data Structures 7
     Further Reading 8
   Estimates of Location 8
     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 17
     Example: Variability Estimates of State Population 18
     Further Reading 19
   Exploring the Data Distribution 19
     Percentiles and Boxplots 20
     Frequency Table and Histograms 21
     Density Estimates 24
     Further Reading 26
   Exploring Binary and Categorical Data 26
     Mode 28
     Expected Value 28
     Further Reading 29
2. Data and Sampling Distributions .......................................................... 43
   Random Sampling and Sample Bias  
      Bias  
      Random Selection  
      Size versus Quality: When Does Size Matter?  
      Sample Mean versus Population Mean  
      Further Reading  
   Selection Bias  
      Regression to the Mean  
      Further Reading  
   Sampling Distribution of a Statistic  
      Central Limit Theorem  
      Standard Error  
      Further Reading  
   The Bootstrap  
      Resampling versus Bootstrapping  
      Further Reading  
   Confidence Intervals  
      Further Reading  
   Normal Distribution  
      Standard Normal and QQ-Plots  
   Long-Tailed Distributions  
      Further Reading  
   Student's t-Distribution  
      Further Reading  
   Binomial Distribution  
      Further Reading  
   Poisson and Related Distributions  
      Poisson Distributions  
      Exponential Distribution  
      Estimating the Failure Rate
### 3. Statistical Experiments and Significance Testing. ........................................ 79

- **A/B Testing**
  - Why Have a Control Group?  
  - Why Just A/B? Why Not C, D…?  
  - For Further Reading 84

- **Hypothesis Tests**
  - The Null Hypothesis 86
  - Alternative Hypothesis 86
  - One-Way, Two-Way Hypothesis Test 87
  - Further Reading 88

- **Resampling**
  - Permutation Test 88
  - Example: Web Stickiness 89
  - Exhaustive and Bootstrap Permutation Test 92
  - Permutation Tests: The Bottom Line for Data Science 93
  - For Further Reading 93

- **Statistical Significance and P-Values**
  - P-Value 96
  - Alpha 96
  - Type 1 and Type 2 Errors 98
  - Data Science and P-Values 98
  - Further Reading 99

- **t-Tests**
  - Further Reading 101

- **Multiple Testing**
  - Further Reading 104

- **Degrees of Freedom**
  - Further Reading 106

- **ANOVA**
  - F-Statistic 109
  - Two-Way ANOVA 110
  - Further Reading 111

- **Chi-Square Test**
  - Chi-Square Test: A Resampling Approach 112
  - Chi-Squared Test: Statistical Theory 114
  - Fisher’s Exact Test 115
  - Relevance for Data Science 117
  - Further Reading 118
4. **Regression and Prediction** ................................................................. 127
   Simple Linear Regression 127
      The Regression Equation 129
      Fitted Values and Residuals 131
      Least Squares 132
      Prediction versus Explanation (Profiling) 133
      Further Reading 134
   Multiple Linear Regression 134
      Example: King County Housing Data 135
      Assessing the Model 136
      Cross-Validation 138
      Model Selection and Stepwise Regression 139
      Weighted Regression 141
   Prediction Using Regression 142
      The Dangers of Extrapolation 143
      Confidence and Prediction Intervals 143
   Factor Variables in Regression 145
      Dummy Variables Representation 145
      Factor Variables with Many Levels 147
      Ordered Factor Variables 149
   Interpreting the Regression Equation 150
      Correlated Predictors 150
      Multicollinearity 151
      Confounding Variables 152
      Interactions and Main Effects 153
   Testing the Assumptions: Regression Diagnostics 155
      Outliers 156
      Influential Values 158
      Heteroskedasticity, Non-Normality and Correlated Errors 161
      Partial Residual Plots and Nonlinearity 164
   Polynomial and Spline Regression 166
      Polynomial 167
      Splines 168
      Generalized Additive Models 170
      Further Reading 172
5. Classification ................................................................. 173
   Naive Bayes 174
   Why Exact Bayesian Classification Is Impractical 175
   The Naive Solution 176
   Numeric Predictor Variables 178
   Further Reading 178
   Discriminant Analysis 179
   Covariance Matrix 180
   Fisher's Linear Discriminant 180
   A Simple Example 181
   Further Reading 183
   Logistic Regression 184
   Logistic Response Function and Logit 184
   Logistic Regression and the GLM 186
   Generalized Linear Models 187
   Predicted Values from Logistic Regression 188
   Interpreting the Coefficients and Odds Ratios 188
   Linear and Logistic Regression: Similarities and Differences 190
   Assessing the Model 191
   Further Reading 194
   Evaluating Classification Models 194
   Confusion Matrix 195
   The Rare Class Problem 196
   Precision, Recall, and Specificity 197
   ROC Curve 198
   AUC 200
   Lift 201
   Further Reading 202
   Strategies for Imbalanced Data 203
   Undersampling 204
   Oversampling and Up/Down Weighting 204
   Data Generation 205
   Cost-Based Classification 206
   Exploring the Predictions 206
   Further Reading 208
   Summary 208

   K-Nearest Neighbors 210
   A Small Example: Predicting Loan Default 211
7. Unsupervised Learning

Principal Components Analysis
- A Simple Example
- Computing the Principal Components
- Interpreting Principal Components

Further Reading

K-Means Clustering
- A Simple Example
- K-Means Algorithm
- Interpreting the Clusters
- Selecting the Number of Clusters

Hierarchical Clustering
- A Simple Example
- The Dendrogram
- The Agglomerative Algorithm
- Measures of Dissimilarity
- Model-Based Clustering

Further Reading