

Data Science & **Big Data Analytics**

Discovering, Analyzing, Visualizing
and Presenting Data

EMC Education Services

WILEY

Contents

<i>Introduction</i>	xvii
---------------------------	------

Chapter 1 • Introduction to Big Data Analytics 1

1.1 Big Data Overview	2
1.1.1 Data Structures	5
1.1.2 Analyst Perspective on Data Repositories	9
1.2 State of the Practice in Analytics	11
1.2.1 BI Versus Data Science	12
1.2.2 Current Analytical Architecture	13
1.2.3 Drivers of Big Data	15
1.2.4 Emerging Big Data Ecosystem and a New Approach to Analytics	16
1.3 Key Roles for the New Big Data Ecosystem	19
1.4 Examples of Big Data Analytics	22
Summary	23
Exercises	23
Bibliography	24

Chapter 2 • Data Analytics Lifecycle..... 25

2.1 Data Analytics Lifecycle Overview	26
2.1.1 Key Roles for a Successful Analytics Project	26
2.1.2 Background and Overview of Data Analytics Lifecycle	28
2.2 Phase 1: Discovery	30
2.2.1 Learning the Business Domain	30
2.2.2 Resources	31
2.2.3 Framing the Problem	32
2.2.4 Identifying Key Stakeholders	33
2.2.5 Interviewing the Analytics Sponsor	33
2.2.6 Developing Initial Hypotheses	35
2.2.7 Identifying Potential Data Sources	35
2.3 Phase 2: Data Preparation	36
2.3.1 Preparing the Analytic Sandbox	37
2.3.2 Performing ETLT	38
2.3.3 Learning About the Data	39
2.3.4 Data Conditioning	40
2.3.5 Survey and Visualize	41
2.3.6 Common Tools for the Data Preparation Phase	42
2.4 Phase 3: Model Planning	42
2.4.1 Data Exploration and Variable Selection	44
2.4.2 Model Selection	45
2.4.3 Common Tools for the Model Planning Phase	45

2.5 Phase 4: Model Building	46
2.5.1 <i>Common Tools for the Model Building Phase</i>	48
2.6 Phase 5: Communicate Results	49
2.7 Phase 6: Operationalize	50
2.8 Case Study: Global Innovation Network and Analysis (GINA)	53
2.8.1 <i>Phase 1: Discovery</i>	54
2.8.2 <i>Phase 2: Data Preparation</i>	55
2.8.3 <i>Phase 3: Model Planning</i>	56
2.8.4 <i>Phase 4: Model Building</i>	56
2.8.5 <i>Phase 5: Communicate Results</i>	58
2.8.6 <i>Phase 6: Operationalize</i>	59
Summary	60
Exercises	61
Bibliography	61
Chapter 3 • Review of Basic Data Analytic Methods Using R	63
3.1 Introduction to R	64
3.1.1 <i>R Graphical User Interfaces</i>	67
3.1.2 <i>Data Import and Export</i>	69
3.1.3 <i>Attribute and Data Types</i>	71
3.1.4 <i>Descriptive Statistics</i>	79
3.2 Exploratory Data Analysis	80
3.2.1 <i>Visualization Before Analysis</i>	82
3.2.2 <i>Dirty Data</i>	85
3.2.3 <i>Visualizing a Single Variable</i>	88
3.2.4 <i>Examining Multiple Variables</i>	91
3.2.5 <i>Data Exploration Versus Presentation</i>	99
3.3 Statistical Methods for Evaluation	101
3.3.1 <i>Hypothesis Testing</i>	102
3.3.2 <i>Difference of Means</i>	104
3.3.3 <i>Wilcoxon Rank-Sum Test</i>	108
3.3.4 <i>Type I and Type II Errors</i>	109
3.3.5 <i>Power and Sample Size</i>	110
3.3.6 <i>ANOVA</i>	110
Summary	114
Exercises	114
Bibliography	115
Chapter 4 • Advanced Analytical Theory and Methods: Clustering	117
4.1 Overview of Clustering	118
4.2 K-means	118
4.2.1 <i>Use Cases</i>	119
4.2.2 <i>Overview of the Method</i>	120
4.2.3 <i>Determining the Number of Clusters</i>	123
4.2.4 <i>Diagnostics</i>	128

4.2.5 <i>Reasons to Choose and Cautions</i>	130
4.3 Additional Algorithms	134
Summary	135
Exercises	135
Bibliography	136
Chapter 5 • Advanced Analytical Theory and Methods: Association Rules	137
5.1 Overview	138
5.2 Apriori Algorithm	140
5.3 Evaluation of Candidate Rules	141
5.4 Applications of Association Rules	143
5.5 An Example: Transactions in a Grocery Store	143
5.5.1 <i>The Groceries Dataset</i>	144
5.5.2 <i>Frequent Itemset Generation</i>	146
5.5.3 <i>Rule Generation and Visualization</i>	152
5.6 Validation and Testing	157
5.7 Diagnostics	158
Summary	158
Exercises	159
Bibliography	160
Chapter 6 • Advanced Analytical Theory and Methods: Regression	161
6.1 Linear Regression	162
6.1.1 <i>Use Cases</i>	162
6.1.2 <i>Model Description</i>	163
6.1.3 <i>Diagnostics</i>	173
6.2 Logistic Regression	178
6.2.1 <i>Use Cases</i>	179
6.2.2 <i>Model Description</i>	179
6.2.3 <i>Diagnostics</i>	181
6.3 Reasons to Choose and Cautions	188
6.4 Additional Regression Models	189
Summary	190
Exercises	190
Chapter 7 • Advanced Analytical Theory and Methods: Classification	191
7.1 Decision Trees	192
7.1.1 <i>Overview of a Decision Tree</i>	193
7.1.2 <i>The General Algorithm</i>	197
7.1.3 <i>Decision Tree Algorithms</i>	203
7.1.4 <i>Evaluating a Decision Tree</i>	204
7.1.5 <i>Decision Trees in R</i>	206
7.2 Naïve Bayes	211
7.2.1 <i>Bayes' Theorem</i>	212
7.2.2 <i>Naïve Bayes Classifier</i>	214

7.2.3 Smoothing	217
7.2.4 Diagnostics.....	217
7.2.5 Naïve Bayes in R	218
7.3 Diagnostics of Classifiers	224
7.4 Additional Classification Methods.....	228
Summary	229
Exercises	230
Bibliography.....	231
Chapter 8 • Advanced Analytical Theory and Methods: Time Series Analysis	233
8.1 Overview of Time Series Analysis	234
8.1.1 Box-Jenkins Methodology.....	235
8.2 ARIMA Model.....	236
8.2.1 Autocorrelation Function (ACF).....	236
8.2.2 Autoregressive Models.....	238
8.2.3 Moving Average Models	239
8.2.4 ARMA and ARIMA Models.....	241
8.2.5 Building and Evaluating an ARIMA Model	244
8.2.6 Reasons to Choose and Cautions	252
8.3 Additional Methods.....	253
Summary	254
Exercises	254
Chapter 9 • Advanced Analytical Theory and Methods: Text Analysis.....	255
9.1 Text Analysis Steps.....	257
9.2 A Text Analysis Example.....	259
9.3 Collecting Raw Text.....	260
9.4 Representing Text	264
9.5 Term Frequency—Inverse Document Frequency (TFIDF)	269
9.6 Categorizing Documents by Topics	274
9.7 Determining Sentiments	277
9.8 Gaining Insights	283
Summary	290
Exercises	290
Bibliography.....	291
Chapter 10 • Advanced Analytics—Technology and Tools: MapReduce and Hadoop.....	295
10.1 Analytics for Unstructured Data	296
10.1.1 Use Cases.....	296
10.1.2 MapReduce	298
10.1.3 Apache Hadoop	300
10.2 The Hadoop Ecosystem	306
10.2.1 Pig.....	306
10.2.2 Hive	308
10.2.3 HBase.....	311
10.2.4 Mahout.....	319

10.3 NoSQL	322
Summary	323
Exercises	324
Bibliography	324
Chapter 11 • Advanced Analytics—Technology and Tools: In-Database Analytics	327
11.1 SQL Essentials	328
11.1.1 Joins	330
11.1.2 Set Operations	332
11.1.3 Grouping Extensions	334
11.2 In-Database Text Analysis	338
11.3 Advanced SQL	343
11.3.1 Window Functions	343
11.3.2 User-Defined Functions and Aggregates	347
11.3.3 Ordered Aggregates	351
11.3.4 MADlib	352
Summary	356
Exercises	356
Bibliography	357
Chapter 12 • The Endgame, or Putting It All Together	359
12.1 Communicating and Operationalizing an Analytics Project	360
12.2 Creating the Final Deliverables	362
12.2.1 Developing Core Material for Multiple Audiences	364
12.2.2 Project Goals	365
12.2.3 Main Findings	367
12.2.4 Approach	369
12.2.5 Model Description	371
12.2.6 Key Points Supported with Data	372
12.2.7 Model Details	372
12.2.8 Recommendations	374
12.2.9 Additional Tips on Final Presentation	375
12.2.10 Providing Technical Specifications and Code	376
12.3 Data Visualization Basics	377
12.3.1 Key Points Supported with Data	378
12.3.2 Evolution of a Graph	380
12.3.3 Common Representation Methods	386
12.3.4 How to Clean Up a Graphic	387
12.3.5 Additional Considerations	392
Summary	393
Exercises	394
References and Further Reading	394
Bibliography	394
<i>Index</i>	397