

# **The Essentials of Data Science**

**Knowledge Discovery Using R**

**Graham J. Williams**



**CRC Press**

Taylor & Francis Group  
Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group, an **Informa** business  
A CHAPMAN & HALL BOOK

---

# *Contents*

---

Preface	ix
List of Figures	xvii
List of Tables	xix
<b>1 Data Science</b>	<b>1</b>
1.1 Exercises . . . . .	12
<b>2 Introducing R</b>	<b>13</b>
2.1 Tooling For R Programming . . . . .	16
2.2 Packages and Libraries . . . . .	22
2.3 Functions, Commands and Operators . . . . .	27
2.4 Pipes . . . . .	31
2.5 Getting Help . . . . .	40
2.6 Exercises . . . . .	41
<b>3 Data Wrangling</b>	<b>43</b>
3.1 Data Ingestion . . . . .	44
3.2 Data Review . . . . .	51
3.3 Data Cleaning . . . . .	54
3.4 Variable Roles . . . . .	63
3.5 Feature Selection . . . . .	66
3.6 Missing Data . . . . .	77
3.7 Feature Creation . . . . .	80
3.8 Preparing the Metadata . . . . .	85
3.9 Preparing for Model Building . . . . .	88
3.10 Save the Dataset . . . . .	92
3.11 A Template for Data Preparation . . . . .	94
3.12 Exercises . . . . .	95

<b>4</b>	<b>Visualising Data</b>	<b>97</b>
4.1	Preparing the Dataset . . . . .	98
4.2	Scatter Plot . . . . .	100
4.3	Bar Chart . . . . .	102
4.4	Saving Plots to File . . . . .	103
4.5	Adding Spice to the Bar Chart . . . . .	103
4.6	Alternative Bar Charts . . . . .	107
4.7	Box Plots . . . . .	111
4.8	Exercises . . . . .	118
<b>5</b>	<b>Case Study: Australian Ports</b>	<b>119</b>
5.1	Data Ingestion . . . . .	120
5.2	Bar Chart: Value/Weight of Sea Trade . . . . .	123
5.3	Scatter Plot: Throughput versus Annual Growth . . . . .	130
5.4	Combined Plots: Port Calls . . . . .	138
5.5	Further Plots . . . . .	141
5.6	Exercises . . . . .	147
<b>6</b>	<b>Case Study: Web Analytics</b>	<b>149</b>
6.1	Sourcing Data from CKAN . . . . .	150
6.2	Browser Data . . . . .	155
6.3	Entry Pages . . . . .	166
6.4	Exercises . . . . .	174
<b>7</b>	<b>A Pattern for Predictive Modelling</b>	<b>175</b>
7.1	Loading the Dataset . . . . .	177
7.2	Building a Decision Tree Model . . . . .	180
7.3	Model Performance . . . . .	185
7.4	Evaluating Model Generality . . . . .	193
7.5	Model Tuning . . . . .	201
7.6	Comparison of Performance Measures . . . . .	209
7.7	Save the Model to File . . . . .	210
7.8	A Template for Predictive Modelling . . . . .	212
7.9	Exercises . . . . .	212
<b>8</b>	<b>Ensemble of Predictive Models</b>	<b>215</b>
8.1	Loading the Dataset . . . . .	216
8.2	Random Forest . . . . .	217

8.3	Extreme Gradient Boosting . . . . .	227
8.4	Exercises . . . . .	239
<b>9</b>	<b>Writing Functions in R</b>	<b>241</b>
9.1	Model Evaluation . . . . .	242
9.2	Creating a Function . . . . .	243
9.3	Function for ROC Curves . . . . .	254
9.4	Exercises . . . . .	256
<b>10</b>	<b>Literate Data Science</b>	<b>257</b>
10.1	Basic L <sup>A</sup> T <sub>E</sub> X Template . . . . .	259
10.2	A Template for our Narrative . . . . .	260
10.3	Including R Commands . . . . .	263
10.4	Inline R Code . . . . .	265
10.5	Formatting Tables Using Kable . . . . .	266
10.6	Formatting Tables Using XTable . . . . .	270
10.7	Including Figures . . . . .	276
10.8	Add a Caption and Label . . . . .	281
10.9	Knitr Options . . . . .	282
10.10	Exercises . . . . .	283
<b>11</b>	<b>R with Style</b>	<b>285</b>
11.1	Why We Should Care . . . . .	285
11.2	Naming . . . . .	287
11.3	Comments . . . . .	291
11.4	Layout . . . . .	292
11.5	Functions . . . . .	298
11.6	Assignment . . . . .	302
11.7	Miscellaneous . . . . .	304
11.8	Exercises . . . . .	305
	<b>Bibliography</b>	<b>307</b>
	<b>Index</b>	<b>313</b>