

# Table of Contents

<b>Preface</b>	<b>1</b>
<b>Chapter 1: Setting the Scene</b>	<b>7</b>
A process framework	8
Data volume and velocity	10
Data variety, formats, and meanings	11
Missing data	12
Cleaning data	12
Visualizing data	13
Resource constraints	13
Terminology	14
Accompanying material	15
Summary	16
<b>Chapter 2: Loading Data</b>	<b>17</b>
<b>Reading files</b>	<b>17</b>
Alternative delimiters	20
Reading complete lines	21
Reading large numbers of attributes	21
Splitting files into smaller pieces	23
<b>Databases</b>	<b>25</b>
The Read Database operator	25
Large datasets	27
<b>Using macros</b>	<b>27</b>
<b>Summary</b>	<b>28</b>

<b>Chapter 3: Visualizing Data</b>	<b>29</b>
<b>Getting started</b>	<b>29</b>
<b>Statistical summaries</b>	<b>30</b>
<b>Relationships between attributes</b>	<b>32</b>
Scatter plots	32
Scatter 3D color	34
Parallel and deviation	35
Quartile color	38
<b>Time series data</b>	<b>39</b>
Plotting series	39
Using the survey plotter	42
<b>Relations between examples</b>	<b>43</b>
Using histograms	44
Using block plots	45
<b>Summary</b>	<b>47</b>
<b>Chapter 4: Parsing and Converting Attributes</b>	<b>49</b>
<b>Generating attributes</b>	<b>50</b>
Date functions	51
Regular expression functions	53
Generating extracts	54
Regular expressions	54
XPath	57
<b>Renaming attributes</b>	<b>59</b>
Searching and replacing attribute values	59
Using the Map operator	59
Using the Replace operator	60
Using the Replace (Dictionary) operator	60
<b>Summary</b>	<b>62</b>
<b>Chapter 5: Outliers</b>	<b>63</b>
<b>Manual inspection</b>	<b>63</b>
Increasing the data volume	68
Rules for handling outliers	68
<b>Automated detection of example outliers</b>	<b>69</b>
The Detect Outlier (Distances) operator	69
The Detect Outlier (Densities) operator	73
The Detect Outlier (LOF) operator	74
The Detect Outliers (COF) operator	75
<b>Summary</b>	<b>76</b>

---

<b>Chapter 6: Missing Values</b>	<b>77</b>
<b>Missing or empty?</b>	<b>77</b>
<b>Types of missing data</b>	<b>78</b>
Missing completely at random	78
Missing at random	78
Not missing at random	79
<b>Categorizing missing data</b>	<b>79</b>
Finding MCAR data	83
Finding MAR data	85
Finding NMAR data	86
A cautionary note	87
<b>Effect of missing data</b>	<b>88</b>
<b>Options for handling missing data</b>	<b>88</b>
Returning to the root cause	89
Ignoring it	89
Manual editing	89
Deletion of examples	90
Deletion of attributes	90
Imputation with single values	90
Modeling	91
<b>Summary</b>	<b>91</b>
<b>Chapter 7: Transforming Data</b>	<b>93</b>
<b>Creating new attributes</b>	<b>94</b>
<b>Aggregation</b>	<b>98</b>
<b>Using pivoting</b>	<b>100</b>
<b>Using de-pivoting</b>	<b>102</b>
<b>Summary</b>	<b>106</b>
<b>Chapter 8: Reducing Data Size</b>	<b>107</b>
<b>Removing examples using sampling</b>	<b>107</b>
<b>Removing attributes</b>	<b>108</b>
Removing useless attributes	109
Weighting attributes	111
Selecting attributes using models	114
<b>Summary</b>	<b>119</b>

<b>Chapter 9: Resource Constraints</b>	<b>121</b>
<b>Measuring and estimating performance</b>	<b>121</b>
Measuring performance	122
<b>Adding memory</b>	<b>129</b>
<b>Parallel processing</b>	<b>130</b>
<b>Restructuring processes</b>	<b>131</b>
<b>Summary</b>	<b>131</b>
<b>Chapter 10: Debugging</b>	<b>133</b>
<b>Breakpoints in RapidMiner Studio</b>	<b>133</b>
<b>Logging data in RapidMiner Studio</b>	<b>134</b>
<b>RapidMiner Studio console printing</b>	<b>135</b>
<b>Groovy scripts</b>	<b>136</b>
Outputting macros example	137
Console logging with Groovy	137
<b>Regex tools</b>	<b>138</b>
<b>Using XPath effectively</b>	<b>138</b>
<b>Summary</b>	<b>139</b>
<b>Chapter 11: Taking Stock</b>	<b>141</b>
<b>Exploring new techniques</b>	<b>142</b>
Time series	142
Web mining	142
Using R	142
Java or Groovy	142
Third-party components	143
RapidMiner Server	143
<b>Where to go next</b>	<b>143</b>
<b>Index</b>	<b>145</b>

---