

Contents

13 Non-Technical Aspects of Quantitative and Data Work	827
13.1 First Principles	827
13.1.1 Framework	828
13.1.2 The “Multiple I” Approach	830
13.1.3 Roles and Responsibilities	831
13.1.4 Analysis Cheatsheet	833
13.2 Project Life Cycle	834
13.2.1 Marketing	834
13.2.2 Initial Contact	836
13.2.3 Client Meetings	837
13.2.4 Assembling the Team	838
13.2.5 Team Meetings	838
13.2.6 Proposal	839
13.2.7 Contracting and IP	842
13.2.8 Project Planning	843
13.2.9 Information Gathering	845
13.2.10 Quantitative Analysis	846
13.2.11 Interpreting the Results	846
13.2.12 Reporting and Deliverables	847
13.2.13 Invoicing	848
13.2.14 Closing the File	849
13.3 Lessons Learned	849
13.3.1 About Clients	849
13.3.2 About Consultants	853
13.4 Business Development	856
13.4.1 Basics	856
13.4.2 Clients and Choices	856
13.4.3 Building Trust	857
13.4.4 Improving Trust	860
13.5 Technical Writing	862
13.5.1 Basics	862
13.5.2 Components	864
13.5.3 Traits	868
13.5.4 Examples	872
13.6 A Conversation With ...	877
13.7 Exercises	878
Chapter References	880
14 Data Science Basics	881
14.1 Introduction	881
14.1.1 What is Data?	881
14.1.2 Objects and Attributes	883
14.1.3 Data in the News	883
14.1.4 Analog vs Digital Data	884

14.2	Conceptual Frameworks	885
14.2.1	Three Modeling Strategies	886
14.2.2	Information Gathering	887
14.2.3	Cognitive Biases	892
14.3	Data Ethics	893
14.3.1	The Need for Ethics	894
14.3.2	What Is/Are Ethics?	894
14.3.3	Ethics and Data Science	895
14.3.4	Guiding Principles	896
14.3.5	The Good, the Bad, the Ugly	898
14.4	Analytics Workflows	898
14.4.1	The “Analytical” Method	898
14.4.2	Collection and Processing	901
14.4.3	Model Assessment	902
14.4.4	Automated Data Pipeline	903
14.5	Getting Insight From Data	903
14.5.1	Asking the Right Questions	905
14.5.2	Structure and Organization	909
14.5.3	Basic Analysis Techniques	919
14.5.4	Common Procedures in R	927
14.5.5	Quantitative Methods	940
14.5.6	Quantitative Fallacies	944
14.6	Exercises	946
	Chapter References	948
15	Data Preparation	951
15.1	Introduction	951
15.2	General Principles	952
15.2.1	Data Cleaning Approaches	952
15.2.2	Pros and Cons	952
15.2.3	Tools and Methods	953
15.3	Data Quality	954
15.3.1	Common Error Sources	955
15.3.2	Detecting Invalid Entries	955
15.4	Missing Values	957
15.4.1	Missing Value Mechanisms	957
15.4.2	Imputation Methods	958
15.4.3	Multiple Imputation	965
15.5	Anomalous Observations	966
15.5.1	Anomaly Detection	967
15.5.2	Outlier Tests	967
15.5.3	Visual Outlier Detection	970
15.6	Data Transformations	972
15.6.1	Common Transformations	973
15.6.2	Box-Cox Transformations	975
15.6.3	Scaling	979
15.6.4	Discretizing	979
15.6.5	Creating Variables	980
15.7	Example: Algae Blooms	980
15.7.1	Problem Description	980
15.7.2	Loading the Data	981

15.7.3	Summary & Visualization	982
15.7.4	Data Cleaning	993
15.7.5	Principal Components	997
15.8	Exercises	999
	Chapter References	1000
16	Web Scraping and Automatic Data Collection	1001
16.1	Data Analysis & Scraping	1001
16.1.1	Why Web Scraping?	1003
16.1.2	Web Data Quality	1003
16.1.3	Ethical Considerations	1005
16.1.4	Decision Process	1007
16.2	Web Technologies Basics	1007
16.2.1	Content Dissemination	1008
16.2.2	Hyper Text Transfer Protocol	1009
16.2.3	Web Content	1010
16.2.4	HTML/XML	1011
16.2.5	Cookies and Other Headers	1011
16.3	Scraping Toolbox	1012
16.3.1	Developer Tools	1012
16.3.2	XPath	1013
16.3.3	Regular Expressions	1023
16.3.4	BeautifulSoup	1027
16.3.5	Selenium	1033
16.3.6	APIs	1033
16.3.7	Specialized Uses	1034
16.4	Examples	1034
16.4.1	Wikipedia	1034
16.4.2	Weather Data	1041
16.4.3	CFL Play-by-Play	1049
16.4.4	Bad HTML	1056
16.4.5	Extracting Text from PDF	1057
16.4.6	YouTube Video Titles	1059
16.5	Exercises	1063
	Chapter References	1064
17	Data Engineering and Data Management	1065
17.1	Background and Context	1065
17.2	Data Engineering	1067
17.2.1	Data Pipelines	1068
17.2.2	Automatic Deployment	1073
17.2.3	Scheduled Pipelines	1075
17.2.4	Data Engineering Tools	1077
17.3	Data Management	1079
17.3.1	Databases	1079
17.3.2	Database Modeling	1082
17.3.3	Data Storage	1084
17.4	Reporting and Deployment	1086
17.4.1	Reports and Products	1086
17.4.2	Cloud vs. On-Premise	1087
	Chapter References	1088

18 Data Exploration and Data Visualization	1089
18.1 Data and Charts	1089
18.1.1 Pre-Analysis Uses	1090
18.1.2 Presenting Results	1090
18.1.3 Multivariate Elements	1091
18.1.4 Visualization Catalogue	1096
18.1.5 Accessibility	1099
18.2 Analytical Design	1099
18.2.1 Comparisons	1100
18.2.2 Mechanism/Explanation	1102
18.2.3 Multivariate Analysis	1104
18.2.4 Integration of Evidence	1106
18.2.5 Documentation	1107
18.2.6 Content First	1110
18.3 Dashboards	1111
18.3.1 Dashboard Fundamentals	1111
18.3.2 Dashboard Structure	1113
18.3.3 Dashboard Design	1114
18.3.4 Examples	1115
18.4 Exercises	1116
Chapter References	1118

List of Figures

13.1 A data science team in action	831
13.2 Avengers, assemble!	838
13.3 The communication continuum	863
13.4 The 5 components of TW; comparison with essay writing	865
13.5 Build-a-shark LEGO instructions	866
13.6 Changing a flat tire	868
13.7 Accessible description	871
14.1 A schematic diagram of systems thinking as it applies to a general problem	890
14.2 A conceptual model of the 'free software' system	890
14.3 UML diagram and ER conceptual map	891
14.4 The analytic workflow	899
14.5 Theoretical and corrupted CRISP-DM processes	900
14.6 The four analytical buckets	905
14.7 Different data cultures and terms	910
14.8 ER model diagram crow's foot relationship symbols cheat sheet	915
14.9 An implemented automated pipeline, with stages and transitions	918
14.10 AFM image of 1,5,9-trioxo-13-azatriangulene with chemical structure	919
14.11 Analysis and pattern-reveal through visualization	926
14.12 The trousers of classification	940
14.13 S&P stock price index, earnings, dividends, and interest rates (1871-2009)	942
14.14 Sales for 3 different products, measured in years, quarters, weeks	943

14.15	Illustration of Simpson's paradox	945
15.1	Data cleaning bingo card	953
15.2	Accuracy as bias, precision as standard error	954
15.3	An illustration of heaping behaviour	955
15.4	Imputation results in the grades data frame	964
15.5	Tukey's boxplot test	968
15.6	Summary visualisations for a plant dataset	970
15.7	Visualisations for a service point dataset	971
15.8	Frequency of the appendage lengths	972
15.9	Illustration of the curse of dimensionality	973
15.10	Various data transformations for a subset of the BUPA liver disease dataset	975
16.1	<i>Robots Exclusion Protocol</i> files	1005
16.2	Etiquette flow diagram for web scraping	1006
16.3	NHL's Atlantic Division standings on 20-Mar-2018	1007
16.4	NHL's Atlantic Division standings under the hood	1008
16.5	Comparison between HTML, XML, and JSON code	1009
16.6	Schematics of HTTP and AJAX requests	1010
16.7	Inspecting a website's elements using Developer Tools	1012
16.8	A simple HTML document, rendered in a browser	1013
16.9	HTML document tree for <code>fortunes.html</code>	1015
16.10	Generic node relations	1017
16.11	7-day forecast for Ottawa, ON	1042
16.12	7-day forecast for Ottawa, ON: <code>class=div-table</code>	1043
16.13	7-day forecast for Ottawa, ON: <code>class=div-column</code>	1043
16.14	7-day forecast for Ottawa, ON: <code>class=div-row1</code>	1045
16.15	7-day forecast for Ottawa, ON: <code>class=div-row2</code>	1045
16.16	CFL 2016 schedule and results	1051
16.17	CFL 2016 schedule and results: heading <code>collapsible-header</code>	1052
16.18	Play-by-play data for a CFL game	1054
16.19	Play-by-play data for a CFL game: <code>playbyplay-tab</code>	1054
16.20	Introduction to Quantitative Consulting YouTube playlist	1061
17.1	Diagram of a conceptual data pipeline	1069
17.2	Data visualization pipeline with component options	1071
17.3	An open-source data analysis pipeline	1077
17.4	A common data analysis pipeline	1078
18.1	Data visualization suggestions, by type of question	1091
18.2	Minimalist NASA CM1 scatterplot	1092
18.3	Deluxe NASA CM1 scatterplot	1092
18.4	2012 Health and Wealth of Nations (Gapminder)	1093
18.5	Choropleths of US elevation	1094
18.6	Network diagram of European language lexical distance	1095
18.7	Physical visualization of people's demographics	1096
18.8	Histogram of reported weekly work hours	1096
18.9	Classification scheme for the kyphosis dataset	1097
18.10	Estimated average project effort overlaid over product complexity, programmer capability, and product count in NASA's COCOMO dataset	1097
18.11	Trend, seasonality, shifts of a supply chain metric	1098
18.12	Diagnosis network around COPD in the Danish Medical Dataset	1098

18.13	Classification of two categories in an artificial dataset	1098
18.14	Hertzsprung-Russell diagram of stellar evolution	1099
18.15	Comparisons in the Gapminder chart: country-to-country	1101
18.16	Comparisons in the Gapminder chart: country-to-country overlap	1101
18.17	Comparisons in the Gapminder chart: country-to-average	1102
18.18	Approximate line of best fit for the Gapminder chart	1103
18.19	Close-up, bottom right quadrant	1103
18.20	Potential outliers in the Gapminder chart	1104
18.21	Potential clusters in the Gapminder chart	1105
18.22	Non-integrated Gapminder chart	1107
18.23	Legend inset for the Gapminder chart	1109
18.24	2013 Health and Wealth of Nations	1111
18.25	Exploratory dashboard for the Global Cities Index dataset	1113
18.26	Exploratory dashboard for the NHL draft class of 2015	1114
18.27	Anonymous 'ugly' dashboards	1115
18.28	'Zen' dashboard I	1119
18.29	'Zen' dashboard II	1120

List of Tables

13.1	An example of weekly time availability for students and professionals	844
13.2	A workplan example	845
15.5	Descriptive statistics for an appendage length dataset	971
16.1	Commonly-used XPath functions	1022