

João Moreira  
André de Carvalho  
Tomáš Horváth

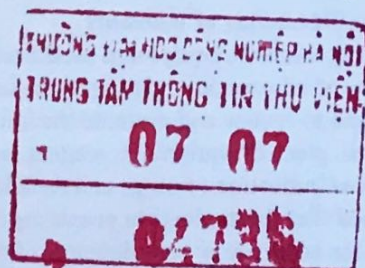
A GENERAL  
INTRODUCTION TO  
**DATA  
ANALYTICS**



**WILEY**

# A General Introduction to Data Analytics

*João Mendes Moreira*  
*University of Porto*



*André C. P. L. F. de Carvalho*  
*University of São Paulo*

*Tomáš Horváth*  
*Eötvös Loránd University in Budapest*  
*Pavol Jozef Šafárik University in Košice*

**WILEY**

This edition first published 2019  
© 2019 John Wiley & Sons, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of João Moreira, André de Carvalho, and Tomáš Horváth to be identified as the author(s) of this work has been asserted in accordance with law.

*Registered Office*

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

*Editorial Office*

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at [www.wiley.com](http://www.wiley.com).

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

*Limit of Liability/Disclaimer of Warranty*

In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

*Library of Congress Cataloging-in-Publication Data*

Names: Moreira, João, 1969– author. | Carvalho, André Carlos Ponce de Leon Ferreira, author. | Horváth, Tomáš, 1976– author.

Title: A general introduction to data analytics / by João Mendes Moreira, André C. P. L. F. de Carvalho, Tomáš Horváth.

Description: Hoboken, NJ : John Wiley & Sons, 2019. | Includes bibliographical references and index. |

Identifiers: LCCN 2017060728 (print) | LCCN 2018005929 (ebook) | ISBN

9781119296256 (pdf) | ISBN 9781119296263 (epub) | ISBN 9781119296249 (cloth)

Subjects: LCSH: Mathematical statistics—Methodology. | Electronic data processing. | Data mining.

Classification: LCC QA276.4 (ebook) | LCC QA276.4 .M664 2018 (print) | DDC 519.50285—dc23

LC record available at <https://lcn.loc.gov/2017060728>

Cover image: © agsandrew/Shutterstock

Cover design by Wiley

Printed in the United States of America.

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

## Contents

Preface	<i>xiii</i>
Acknowledgments	<i>xv</i>
Presentational Conventions	<i>xvii</i>
About the Companion Website	<i>xix</i>

### Part I Introductory Background 1

1	What Can We Do With Data?	3
1.1	Big Data and Data Science	4
1.2	Big Data Architectures	5
1.3	Small Data	6
1.4	What is Data?	7
1.5	A Short Taxonomy of Data Analytics	9
1.6	Examples of Data Use	10
1.6.1	Breast Cancer in Wisconsin	11
1.6.2	Polish Company Insolvency Data	11
1.7	A Project on Data Analytics	12
1.7.1	A Little History on Methodologies for Data Analytics	12
1.7.2	The KDD Process	14
1.7.3	The CRISP-DM Methodology	15
1.8	How this Book is Organized	16
1.9	Who Should Read this Book	18

### Part II Getting Insights from Data 19

2	Descriptive Statistics	21
2.1	Scale Types	22
2.2	Descriptive Univariate Analysis	25
2.2.1	Univariate Frequencies	25

2.2.2	Univariate Data Visualization	27
2.2.3	Univariate Statistics	32
2.2.4	Common Univariate Probability Distributions	38
2.3	Descriptive Bivariate Analysis	40
2.3.1	Two Quantitative Attributes	41
2.3.2	Two Qualitative Attributes, at Least one of them Nominal	45
2.3.3	Two Ordinal Attributes	46
2.4	Final Remarks	47
2.5	Exercises	47
<b>3</b>	<b>Descriptive Multivariate Analysis</b>	<b>49</b>
3.1	Multivariate Frequencies	49
3.2	Multivariate Data Visualization	50
3.3	Multivariate Statistics	59
3.3.1	Location Multivariate Statistics	59
3.3.2	Dispersion Multivariate Statistics	60
3.4	Infographics and Word Clouds	66
3.4.1	Infographics	66
3.4.2	Word Clouds	67
3.5	Final Remarks	67
3.6	Exercises	68
<b>4</b>	<b>Data Quality and Preprocessing</b>	<b>71</b>
4.1	Data Quality	71
4.1.1	Missing Values	72
4.1.2	Redundant Data	74
4.1.3	Inconsistent Data	75
4.1.4	Noisy Data	76
4.1.5	Outliers	77
4.2	Converting to a Different Scale Type	77
4.2.1	Converting Nominal to Relative	78
4.2.2	Converting Ordinal to Relative or Absolute	81
4.2.3	Converting Relative or Absolute to Ordinal or Nominal	82
4.3	Converting to a Different Scale	83
4.4	Data Transformation	85
4.5	Dimensionality Reduction	86
4.5.1	Attribute Aggregation	88
4.5.1.1	Principal Component Analysis	88
4.5.1.2	Independent Component Analysis	91
4.5.1.3	Multidimensional Scaling	91
4.5.2	Attribute Selection	92
4.5.2.1	Filters	92
4.5.2.2	Wrappers	93
4.5.2.3	Embedded	94

4.5.2.4	Search Strategies	95
4.6	Final Remarks	96
4.7	Exercises	96
<b>5</b>	<b>Clustering</b>	<b>99</b>
5.1	Distance Measures	100
5.1.1	Differences between Values of Common Attribute Types	101
5.1.2	Distance Measures for Objects with Quantitative Attributes	103
5.1.3	Distance Measures for Non-conventional Attributes	104
5.2	Clustering Validation	107
5.3	Clustering Techniques	108
5.3.1	K-means	110
5.3.1.1	Centroids and Distance Measures	110
5.3.1.2	How K-means Works	111
5.3.2	DBSCAN	115
5.3.3	Agglomerative Hierarchical Clustering Technique	117
5.3.3.1	Linkage Criterion	119
5.3.3.2	Dendrograms	120
5.4	Final Remarks	122
5.5	Exercises	123
<b>6</b>	<b>Frequent Pattern Mining</b>	<b>125</b>
6.1	Frequent Itemsets	127
6.1.1	Setting the <i>min_sup</i> Threshold	128
6.1.2	Apriori – a Join-based Method	131
6.1.3	Eclat	133
6.1.4	FP-Growth	134
6.1.5	Maximal and Closed Frequent Itemsets	138
6.2	Association Rules	139
6.3	Behind Support and Confidence	142
6.3.1	Cross-support Patterns	143
6.3.2	Lift	144
6.3.3	Simpson's Paradox	145
6.4	Other Types of Pattern	147
6.4.1	Sequential patterns	147
6.4.2	Frequent Sequence Mining	148
6.4.3	Closed and Maximal Sequences	148
6.5	Final Remarks	149
6.6	Exercises	149
<b>7</b>	<b>Cheat Sheet and Project on Descriptive Analytics</b>	<b>151</b>
7.1	Cheat Sheet of Descriptive Analytics	151
7.1.1	On Data Summarization	151

7.1.2	On Clustering	151
7.1.3	On Frequent Pattern Mining	153
7.2	Project on Descriptive Analytics	154
7.2.1	Business Understanding	154
7.2.2	Data Understanding	155
7.2.3	Data Preparation	155
7.2.4	Modeling	157
7.2.5	Evaluation	158
7.2.6	Deployment	158

### Part III Predicting the Unknown 159

<b>8</b>	<b>Regression</b>	161
8.1	Predictive Performance Estimation	164
8.1.1	Generalization	164
8.1.2	Model Validation	165
8.1.3	Predictive Performance Measures for Regression	169
8.2	Finding the Parameters of the Model	171
8.2.1	Linear Regression	171
8.2.1.1	Empirical Error	173
8.2.2	The Bias-variance Trade-off	175
8.2.3	Shrinkage Methods	177
8.2.3.1	Ridge Regression	179
8.2.3.2	Lasso Regression	180
8.2.4	Methods that use Linear Combinations of Attributes	181
8.2.4.1	Principal Components Regression	181
8.2.4.2	Partial Least Squares Regression	182
8.3	Technique and Model Selection	182
8.4	Final Remarks	183
8.5	Exercises	184
<b>9</b>	<b>Classification</b>	187
9.1	Binary Classification	188
9.2	Predictive Performance Measures for Classification	192
9.3	Distance-based Learning Algorithms	199
9.3.1	K-nearest Neighbor Algorithms	199
9.3.2	Case-based Reasoning	202
9.4	Probabilistic Classification Algorithms	203
9.4.1	Logistic Regression Algorithm	205
9.4.2	Naive Bayes Algorithm	207
9.5	Final Remarks	208
9.6	Exercises	210

<b>10</b>	<b>Additional Predictive Methods</b>	<b>211</b>
10.1	Search-based Algorithms	211
10.1.1	Decision Tree Induction Algorithms	212
10.1.2	Decision Trees for Regression	217
10.1.2.1	Model Trees	218
10.1.2.2	Multivariate Adaptive Regression Splines	219
10.2	Optimization-based Algorithms	221
10.2.1	Artificial Neural Networks	222
10.2.1.1	Backpropagation	224
10.2.1.2	Deep Networks and Deep Learning Algorithms	230
10.2.2	Support Vector Machines	233
10.2.2.1	SVM for Regression	237
10.3	Final Remarks	238
10.4	Exercises	239
<b>11</b>	<b>Advanced Predictive Topics</b>	<b>241</b>
11.1	Ensemble Learning	241
11.1.1	Bagging	243
11.1.2	Random Forests	244
11.1.3	AdaBoost	245
11.2	Algorithm Bias	246
11.3	Non-binary Classification Tasks	248
11.3.1	One-class Classification	248
11.3.2	Multi-class Classification	249
11.3.3	Ranking Classification	250
11.3.4	Multi-label Classification	251
11.3.5	Hierarchical Classification	252
11.4	Advanced Data Preparation Techniques for Prediction	253
11.4.1	Imbalanced Data Classification	253
11.4.2	For Incomplete Target Labeling	254
11.4.2.1	Semi-supervised Learning	254
11.4.2.2	Active Learning	255
11.5	Description and Prediction with Supervised Interpretable Techniques	255
11.6	Exercises	256
<b>12</b>	<b>Cheat Sheet and Project on Predictive Analytics</b>	<b>259</b>
12.1	Cheat Sheet on Predictive Analytics	259
12.2	Project on Predictive Analytics	259
12.2.1	Business Understanding	260
12.2.2	Data Understanding	260
12.2.3	Data Preparation	265
12.2.4	Modeling	265
12.2.5	Evaluation	265
12.2.6	Deployment	266

## Part IV Popular Data Analytics Applications 267

### 13 Applications for Text, Web and Social Media 269

- 13.1 Working with Texts 269
  - 13.1.1 Data Acquisition 271
  - 13.1.2 Feature Extraction 271
    - 13.1.2.1 Tokenization 272
    - 13.1.2.2 Stemming 272
    - 13.1.2.3 Conversion to Structured Data 275
    - 13.1.2.4 Is the Bag of Words Enough? 276
  - 13.1.3 Remaining Phases 277
  - 13.1.4 Trends 277
    - 13.1.4.1 Sentiment Analysis 278
    - 13.1.4.2 Web Mining 278
- 13.2 Recommender Systems 278
  - 13.2.1 Feedback 279
  - 13.2.2 Recommendation Tasks 280
  - 13.2.3 Recommendation Techniques 281
    - 13.2.3.1 Knowledge-based Techniques 281
    - 13.2.3.2 Content-based Techniques 282
    - 13.2.3.3 Collaborative Filtering Techniques 282
  - 13.2.4 Final Remarks 289
- 13.3 Social Network Analysis 291
  - 13.3.1 Representing Social Networks 291
  - 13.3.2 Basic Properties of Nodes 294
    - 13.3.2.1 Degree 294
    - 13.3.2.2 Distance 294
    - 13.3.2.3 Closeness 295
    - 13.3.2.4 Betweenness 296
    - 13.3.2.5 Clustering Coefficient 297
  - 13.3.3 Basic and Structural Properties of Networks 297
    - 13.3.3.1 Diameter 297
    - 13.3.3.2 Centralization 297
    - 13.3.3.3 Cliques 299
    - 13.3.3.4 Clustering Coefficient 299
    - 13.3.3.5 Modularity 299
  - 13.3.4 Trends and Final Remarks 299
- 13.4 Exercises 300

### Appendix A: Comprehensive Description of the CRISP-DM Methodology 303

#### References 311

#### Index 315

# A guide to the principles and methods of data analysis that does not require knowledge of statistics or programming

*A General Introduction to Data Analytics* is an essential guide to understand and use data analytics. This book is written using easy-to-understand terms and does not require familiarity with statistics or programming. The authors—noted experts in the field—highlight an explanation of the intuition behind the basic data analytics techniques. The text also contains exercises and illustrative examples.

Thought to be easily accessible to non-experts, the book provides motivation to the necessity of analyzing data. It explains how to visualize and summarize data, and how to find natural groups and frequent patterns in a dataset. The book also explores predictive tasks, be them classification or regression. Finally, the book discusses popular data analytic applications, like mining the web, information retrieval, social network analysis, working with text, and recommender systems. The learning resources offer:

- A guide to the reasoning behind data mining techniques
- A unique illustrative example that extends throughout all the chapters
- Exercises at the end of each chapter and larger projects at the end of parts II and III of the book
- Supplemented with PowerPoint slides available for instructors on a Wiley Book Companion Site

Together with these learning resources, the book can be used in a 13-week course guide, one chapter per course topic.

The book was written in a format that allows the understanding of the main data analytics concepts by non-mathematicians, non-statisticians and non-computer scientists interested in getting an introduction to data science. *A General Introduction to Data Analytics* is a basic guide to data analytics written in highly accessible terms.

**João Mendes Moreira, PhD**, is an assistant professor in the Faculty of Engineering at the University of Porto, Porto, Portugal and is also a researcher in LIAAD-INESC TEC, Porto, Portugal.

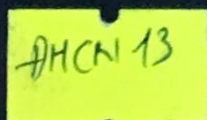
**André de Carvalho, PhD**, is a full professor in the Institute of Mathematics and Computer Science at the University of São Paulo, Brazil.

**Tomáš Horváth, PhD**, is an assistant professor at the Faculty of Informatics of the Eötvös Loránd University in Budapest, Hungary, and is also associated with the Faculty of Science at the Pavol Jozef Šafárik University in Košice, Slovakia.



A companion website with additional resources is available at:  
[www.wiley.com/go/moreira/dataanalytics](http://www.wiley.com/go/moreira/dataanalytics)

Cover Design: Wiley  
Cover Image: © agsandrew/Shutterstock



Also available  
as an e-book

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ  
TRUNG TÂM THÔNG TIN TH



Mã sách: 070704786

ISBN 978-1-119-29624-9



9 781119 296249

**WILEY**