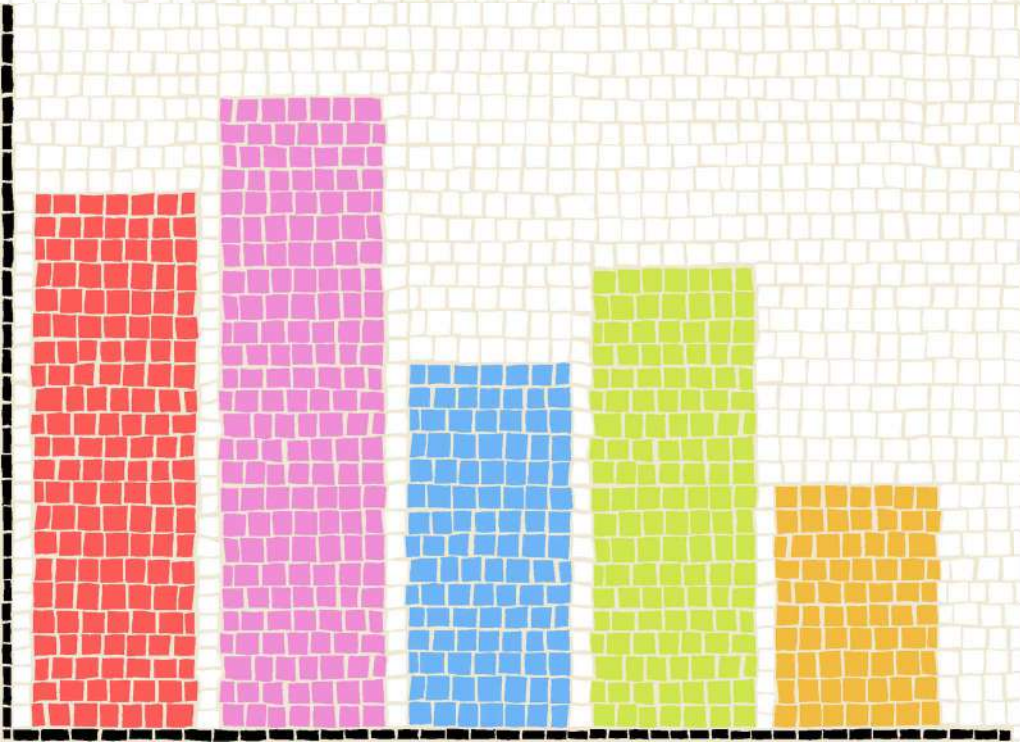


The Art of Data Science

A Guide for Anyone Who Works with Data



Roger D. Peng & Elizabeth Matsui

Contents

1. Data Analysis as Art	1
2. Epicycles of Analysis	4
2.1 Setting the Scene	5
2.2 Epicycle of Analysis	6
2.3 Setting Expectations	8
2.4 Collecting Information	9
2.5 Comparing Expectations to Data	10
2.6 Applying the Epicycle of Analysis Process	11
3. Stating and Refining the Question	16
3.1 Types of Questions	16
3.2 Applying the Epicycle to Stating and Refining Your Question	20
3.3 Characteristics of a Good Question	20
3.4 Translating a Question into a Data Problem	23
3.5 Case Study	26
3.6 Concluding Thoughts	30
4. Exploratory Data Analysis	31
4.1 Exploratory Data Analysis Checklist: A Case Study	33
4.2 Formulate your question	33
4.3 Read in your data	35
4.4 Check the Packaging	36
4.5 Look at the Top and the Bottom of your Data	39

CONTENTS

4.6	ABC: Always be Checking Your “n”s	40
4.7	Validate With at Least One External Data Source	45
4.8	Make a Plot	46
4.9	Try the Easy Solution First	49
4.10	Follow-up Questions	53
5.	Using Models to Explore Your Data	55
5.1	Models as Expectations	57
5.2	Comparing Model Expectations to Reality .	60
5.3	Reacting to Data: Refining Our Expectations	64
5.4	Examining Linear Relationships	67
5.5	When Do We Stop?	73
5.6	Summary	77
6.	Inference: A Primer	78
6.1	Identify the population	78
6.2	Describe the sampling process	79
6.3	Describe a model for the population	79
6.4	A Quick Example	80
6.5	Factors Affecting the Quality of Inference .	84
6.6	Example: Apple Music Usage	86
6.7	Populations Come in Many Forms	89
7.	Formal Modeling	92
7.1	What Are the Goals of Formal Modeling? .	92
7.2	General Framework	93
7.3	Associational Analyses	95
7.4	Prediction Analyses	104
7.5	Summary	111
8.	Inference vs. Prediction: Implications for Modeling Strategy	112
8.1	Air Pollution and Mortality in New York City	113
8.2	Inferring an Association	115
8.3	Predicting the Outcome	121

CONTENTS

8.4 Summary	123
9. Interpreting Your Results	124
9.1 Principles of Interpretation	124
9.2 Case Study: Non-diet Soda Consumption and Body Mass Index	125
10. Communication	144
10.1 Routine communication	144
10.2 The Audience	146
10.3 Content	148
10.4 Style	151
10.5 Attitude	151
11. Concluding Thoughts	153
12. About the Authors	155