

---

# Table of Contents

Preface..... xv

---

## Part I. The Data Science Lifecycle

<b>1. The Data Science Lifecycle.....</b>	<b>3</b>
The Stages of the Lifecycle	3
Examples of the Lifecycle	6
Summary	7
<b>2. Questions and Data Scope.....</b>	<b>9</b>
Big Data and New Opportunities	10
Example: Google Flu Trends	10
Target Population, Access Frame, and Sample	12
Example: What Makes Members of an Online Community Active?	14
Example: Who Will Win the Election?	14
Example: How Do Environmental Hazards Relate to an Individual's Health?	15
Instruments and Protocols	16
Measuring Natural Phenomena	17
Example: What Is the Level of CO <sub>2</sub> in the Air?	18
Accuracy	19
Types of Bias	20
Types of Variation	22
Summary	24

<b>3. Simulation and Data Design.....</b>	<b>27</b>
The Urn Model	28
Sampling Designs	30
Sampling Distribution of a Statistic	32
Simulating the Sampling Distribution	33
Simulation with the Hypergeometric Distribution	35
Example: Simulating Election Poll Bias and Variance	36
The Pennsylvania Urn Model	38
An Urn Model with Bias	40
Conducting Larger Polls	41
Example: Simulating a Randomized Trial for a Vaccine	43
Scope	43
The Urn Model for Random Assignment	44
Example: Measuring Air Quality	46
Summary	49
<b>4. Modeling with Summary Statistics.....</b>	<b>51</b>
The Constant Model	52
Minimizing Loss	54
Mean Absolute Error	55
Mean Squared Error	57
Choosing Loss Functions	59
Summary	60
<b>5. Case Study: Why Is My Bus Always Late?.....</b>	<b>63</b>
Question and Scope	64
Data Wrangling	64
Exploring Bus Times	67
Modeling Wait Times	70
Summary	74

---

## Part II. Rectangular Data

<b>6. Working with Dataframes Using pandas.....</b>	<b>79</b>
Subsetting	80
Data Scope and Question	80
Dataframes and Indices	81
Slicing	83
Filtering Rows	86
Example: How Recently Has Luna Become a Popular Name?	89

Aggregating	91
Basic Group-Aggregate	92
Grouping on Multiple Columns	95
Custom Aggregation Functions	96
Pivoting	98
Joining	100
Inner Joins	101
Left, Right, and Outer Joins	103
Example: Popularity of NYT Name Categories	105
Transforming	107
Apply	107
Example: Popularity of “L” Names	109
The Price of Apply	110
How Are Dataframes Different from Other Data Representations?	111
Dataframes and Spreadsheets	111
Dataframes and Matrices	112
Dataframes and Relations	113
Summary	113
<b>7. Working with Relations Using SQL.....</b>	<b>115</b>
Subsetting	115
SQL Basics: SELECT and FROM	116
What’s a Relation?	117
Slicing	118
Filtering Rows	119
Example: How Recently Has Luna Become a Popular Name?	121
Aggregating	122
Basic Group-Aggregate Using GROUP BY	123
Grouping on Multiple Columns	124
Other Aggregation Functions	125
Joining	126
Inner Joins	127
Left and Right Joins	129
Example: Popularity of NYT Name Categories	130
Transforming and Common Table Expressions	131
SQL Functions	131
Multistep Queries Using a WITH Clause	134
Example: Popularity of “L” Names	134
Summary	135

---

## Part III. Understanding The Data

<b>8. Wrangling Files.....</b>	<b>139</b>
Data Source Examples	140
Drug Abuse Warning Network (DAWN) Survey	140
San Francisco Restaurant Food Safety	140
File Formats	142
Delimited Format	142
Fixed-Width Format	144
Hierarchical Formats	145
Loosely Formatted Text	145
File Encoding	146
File Size	148
The Shell and Command-Line Tools	151
Table Shape and Granularity	155
Granularity of Restaurant Inspections and Violations	156
DAWN Survey Shape and Granularity	158
Summary	161
<b>9. Wrangling Dataframes.....</b>	<b>163</b>
Example: Wrangling CO <sub>2</sub> Measurements from the Mauna Loa Observatory	164
Quality Checks	167
Addressing Missing Data	170
Reshaping the Data Table	171
Quality Checks	172
Quality Based on Scope	172
Quality of Measurements and Recorded Values	173
Quality Across Related Features	174
Quality for Analysis	174
Fixing the Data or Not	175
Missing Values and Records	176
Transformations and Timestamps	178
Transforming Timestamps	179
Piping for Transformations	182
Modifying Structure	183
Example: Wrangling Restaurant Safety Violations	186
Narrowing the Focus	187
Aggregating Violations	188
Extracting Information from Violation Descriptions	190
Summary	193

<b>10. Exploratory Data Analysis.....</b>	<b>195</b>
Feature Types	196
Example: Dog Breeds	198
Transforming Qualitative Features	203
The Importance of Feature Types	206
What to Look For in a Distribution	207
What to Look For in a Relationship	211
Two Quantitative Features	211
One Qualitative and One Quantitative Variable	212
Two Qualitative Features	214
Comparisons in Multivariate Settings	216
Guidelines for Exploration	220
Example: Sale Prices for Houses	221
Understanding Price	222
What Next?	224
Examining Other Features	225
Delving Deeper into Relationships	229
Fixing Location	230
EDA Discoveries	232
Summary	233
<b>11. Data Visualization.....</b>	<b>235</b>
Choosing Scale to Reveal Structure	235
Filling the Data Region	236
Including Zero	237
Revealing Shape Through Transformations	239
Banking to Decipher Relationships	241
Revealing Relationships Through Straightening	242
Smoothing and Aggregating Data	245
Smoothing Techniques to Uncover Shape	245
Smoothing Techniques to Uncover Relationships and Trends	247
Smoothing Techniques Need Tuning	249
Reducing Distributions to Quantiles	250
When Not to Smooth	252
Facilitating Meaningful Comparisons	254
Emphasize the Important Difference	254
Ordering Groups	256
Avoid Stacking	258
Selecting a Color Palette	260
Guidelines for Comparisons in Plots	262
Incorporating the Data Design	263

Data Collected Over Time	263
Observational Studies	265
Unequal Sampling	266
Geographic Data	267
Adding Context	268
Example: 100m Sprint Times	269
Creating Plots Using plotly	270
Figure and Trace Objects	271
Modifying Layout	273
Plotting Functions	274
Annotations	277
Other Tools for Visualization	278
matplotlib	278
Grammar of Graphics	278
Summary	279
<b>12. Case Study: How Accurate Are Air Quality Measurements?.....</b>	<b>281</b>
Question, Design, and Scope	282
Finding Collocated Sensors	284
Wrangling the List of AQS Sites	284
Wrangling the List of PurpleAir Sites	286
Matching AQS and PurpleAir Sensors	288
Wrangling and Cleaning AQS Sensor Data	290
Checking Granularity	291
Removing Unneeded Columns	292
Checking the Validity of Dates	292
Checking the Quality of PM2.5 Measurements	293
Wrangling PurpleAir Sensor Data	294
Checking the Granularity	296
Handling Missing Values	300
Exploring PurpleAir and AQS Measurements	302
Creating a Model to Correct PurpleAir Measurements	308
Summary	310

---

## Part IV. Other Data Sources

<b>13. Working with Text.....</b>	<b>315</b>
Examples of Text and Tasks	316
Convert Text into a Standard Format	316
Extract a Piece of Text to Create a Feature	316

Transform Text into Features	317
Text Analysis	317
String Manipulation	318
Converting Text to a Standard Format with Python String Methods	318
String Methods in pandas	319
Splitting Strings to Extract Pieces of Text	320
Regular Expressions	321
Concatenation of Literals	322
Quantifiers	324
Alternation and Grouping to Create Features	326
Reference Tables	327
Text Analysis	329
Summary	334
<b>14. Data Exchange.....</b>	<b>335</b>
NetCDF Data	336
JSON Data	341
HTTP	345
REST	349
XML, HTML, and XPath	353
Example: Scraping Race Times from Wikipedia	356
XPath	358
Example: Accessing Exchange Rates from the ECB	360
Summary	363

---

## Part V. Linear Modeling

<b>15. Linear Models.....</b>	<b>367</b>
Simple Linear Model	368
Example: A Simple Linear Model for Air Quality	372
Interpreting Linear Models	374
Assessing the Fit	375
Fitting the Simple Linear Model	377
Multiple Linear Model	379
Fitting the Multiple Linear Model	384
Example: Where Is the Land of Opportunity?	388
Explaining Upward Mobility Using Commute Time	389
Relating Upward Mobility Using Multiple Variables	392
Feature Engineering for Numeric Measurements	396
Feature Engineering for Categorical Measurements	400

Summary	407
<b>16. Model Selection.....</b>	<b>409</b>
Overfitting	410
Example: Energy Consumption	410
Train-Test Split	415
Cross-Validation	419
Regularization	424
Model Bias and Variance	425
Summary	429
<b>17. Theory for Inference and Prediction.....</b>	<b>431</b>
Distributions: Population, Empirical, Sampling	431
Basics of Hypothesis Testing	433
Example: A Rank Test to Compare Productivity of Wikipedia Contributors	435
Example: A Test of Proportions for Vaccine Efficacy	439
Bootstrapping for Inference	442
Basics of Confidence Intervals	446
Basics of Prediction Intervals	450
Example: Predicting Bus Lateness	450
Example: Predicting Crab Size	451
Example: Predicting the Incremental Growth of a Crab	453
Probability for Inference and Prediction	455
Formalizing the Theory for Average Rank Statistics	456
General Properties of Random Variables	459
Probability Behind Testing and Intervals	462
Probability Behind Model Selection	465
Summary	467
<b>18. Case Study: How to Weigh a Donkey.....</b>	<b>471</b>
Donkey Study Question and Scope	471
Wrangling and Transforming	472
Exploring	478
Modeling a Donkey's Weight	481
A Loss Function for Prescribing Anesthetics	481
Fitting a Simple Linear Model	482
Fitting a Multiple Linear Model	484
Bringing Qualitative Features into the Model	485
Model Assessment	488
Summary	490



---

## Part VI. Classification

<b>19. Classification.....</b>	<b>495</b>
Example: Wind-Damaged Trees	496
Modeling and Classification	498
A Constant Model	498
Examining the Relationship Between Size and Windthrow	499
Modeling Proportions (and Probabilities)	501
A Logistic Model	502
Log Odds	504
Using a Logistic Curve	505
A Loss Function for the Logistic Model	505
From Probabilities to Classification	509
The Confusion Matrix	511
Precision Versus Recall	512
Summary	515
<b>20. Numerical Optimization.....</b>	<b>517</b>
Gradient Descent Basics	518
Minimizing Huber Loss	520
Convex and Differentiable Loss Functions	522
Variants of Gradient Descent	524
Stochastic Gradient Descent	525
Mini-Batch Gradient Descent	525
Newton's Method	526
Summary	527
<b>21. Case Study: Detecting Fake News.....</b>	<b>529</b>
Question and Scope	530
Obtaining and Wrangling the Data	531
Exploring the Data	535
Exploring the Publishers	536
Exploring Publication Date	538
Exploring Words in Articles	540
Modeling	542
A Single-Word Model	542
Multiple-Word Model	544
Predicting with the tf-idf Transform	546
Summary	549

<b>Additional Material.....</b>	<b>551</b>
<b>Data Sources.....</b>	<b>557</b>
<b>Index.....</b>	<b>561</b>