

Hands-On Data Analysis in R for Finance

Jean-François Collard



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

Contents

List of Figures	xiii
Preface	xix
1 Your Working Environment	1
1.1 RStudio	1
1.2 R Notebooks	2
1.3 Packages	4
1.4 Specialized Packages for Finance	5
2 Reading Data in R	7
2.1 Reading Input (Data) Files	7
2.2 Reading Excel Files	13
2.3 Reading Tables	14
2.4 Packages Come With Datasets	14
2.5 Reading XML Data	15
2.6 JSON	16
2.7 Chapter-End Summary	16
3 Financial Data	17
3.1 Yahoo! Finance	17
3.2 Federal Reserve Economic Data (FRED)	24
3.3 Nasdaq	25
3.4 Other Data Sources	29
4 Introduction to R	31
4.1 Expressions	31
4.2 Creating New Variables	31
4.3 Data Types and Type Conversion	33
4.4 Vectors	34
4.5 Matrices	38
4.6 Lists	45
4.7 Data Frames	52
4.8 Time Series	58
4.9 Data Wrangling	62
4.10 Exercises	65
4.10.1 Formatting	65

4.10.2	Format Conversion	65
4.10.3	Wrangling Using <code>pivot_longer</code>	67
4.10.4	Computing Daily Returns From Daily Prices	67
4.10.5	Histogram of Apple's Daily Returns	68
5	Functions	69
5.1	Calling Existing Functions	69
5.2	Creating New Functions	71
5.3	Function Composition (a.k.a Piping)	72
5.4	Optimization	74
5.5	Manipulating Character Strings	75
5.6	Key Statistics Functions	78
5.7	Empirical Distributions	84
5.8	Chapter-End Summary	88
5.9	Exercises	89
5.9.1	Histogram of Oil Returns	89
5.9.2	ECDF of Oil Prices	90
5.9.3	Peak of Oil Prices	91
5.9.4	Qnorm	91
5.9.5	Returns vs Log Returns	91
5.9.6	Skew and Kurtosis	92
5.9.7	Function to Calculate Returns	92
5.9.8	Risk Limit	92
5.9.9	Probability of Reaching a Profit Target	93
5.9.10	Finding Most Significant Outlier	93
6	Data Transformation	95
6.1	Selecting Rows: Slicing	95
6.2	Group_by	96
6.3	Filter	97
6.4	Arrange	98
6.5	Rename	99
6.6	Mutate	100
6.7	Summarize	102
6.8	Contingency Tables	105
6.9	Aggregate	107
6.10	Chapter-End Summary	110
6.11	Exercises	110
6.11.1	Filtering on Either of Two Conditions	110
6.11.2	Performance by Sector	111
6.11.3	Ordering and Plotting Returns	111
6.11.4	Removing NAs	112
6.11.5	Removing Outliers	112
6.11.6	Deutsche Bank's Long-Term Debt	113

7	Merging Data Sets	115
7.1	Inner Join	116
7.2	Left Join	117
7.3	Right Join	118
7.4	Full Join (a.k.a. Outer Join)	119
7.5	Merging Nasdaq Datasets	121
7.6	Chapter-End Summary	123
7.7	Exercises	123
7.7.1	The Zacks EE Dataset	123
7.7.2	Merging Dividend and Split Data	124
8	Graphing Using Ggplot	127
8.1	The Grammar of Ggplot Commands	127
8.2	Geometric Objects	128
8.3	Separating by Color	132
8.4	Separating by Size	133
8.5	Separating by Shape	133
8.6	Curves of Best Fit	134
8.7	Case Study: The House Price Dataset	144
8.8	Case Study: The Ocean Portfolio	147
8.9	Exercises	153
8.9.1	Change the Marker Shape by Region	153
8.9.2	Change the Marker Color by Price	153
8.9.3	Market Cap by Countries	154
9	Returns and Returns-based Statistics	157
9.1	Single-Period Returns	157
9.2	Multiple Periods	159
9.3	Prices and Adjusted Prices	160
9.4	Returns	163
9.5	Volatility	167
9.6	Sharpe	168
9.7	Drawdowns	170
9.8	Benchmark-Relative Performance and Risk	171
9.9	Rolling Correlations	182
9.10	Normality of Return Distributions	183
9.11	Fitting A Distribution	191
9.12	Are Differences in Returns Significant?	193
9.13	Exercises	196
9.13.1	Verifying GM's and Ford's Returns	196
9.13.2	Computing Monthly Percentage Changes of Oil Prices	197
9.13.3	Comparing Returns and Log>Returns	198
9.13.4	Worst and Best Days for Bitcoin	198
9.13.5	Bull Beta	199

10	Portfolios	201
10.1	Building Portfolios Using Tidyquant	201
10.2	Building Portfolios Using PerformanceAnalytics	205
10.3	Portfolio Optimization	209
10.4	Exercises	221
10.4.1	Correlation Matrix	221
10.4.2	Improving the Portfolio Growth Graph	222
10.4.3	Portfolio of Hedge Funds	224
10.4.4	Larger Search Space	226
11	Modeling Returns & Simulations	227
11.1	Normal and Log-normal Models	227
11.2	Log-normal Model – Multi-period Return	229
11.3	Random Walk	230
11.4	Geometric Random Walk	231
11.5	Toward Simulations	232
11.6	The Multiple Questions Simulations Can Answer	233
11.7	Exercises	238
11.7.1	Probability of a Loss	238
12	Linear and Polynomial Regression	239
12.1	The House Price Dataset	241
12.2	Multi-linear Regression	246
12.3	Collinearity	248
12.4	Variance Inflation Factor	252
12.5	ANOVA	253
12.6	Response Transformation	256
12.7	Linear Regression with Categorical Variables	262
12.8	Polynomial Regression	264
12.9	Exercises	266
12.9.1	Collinearity	266
12.9.2	Order of Independent Variables in Multi-linear Regressions	267
13	Fixed Income	269
13.1	Present Value	270
13.2	Present Value of Coupon Bonds	273
13.3	Exercises	276
13.3.1	Alternative Formula for the Present Value of a Coupon Bond	276
13.3.2	Modified Duration	277
13.3.3	Yield to Maturity	277
14	Principal Component Analysis	279
14.1	Directions of Most Variance	279
14.2	Application to a Full Example	281

- 14.3 How Much Variance is Explained by Each Principal Component? 283
- 14.4 Chapter-End Summary 285
- 14.5 Exercises 286
 - 14.5.1 PCA on Rates 286
 - 14.5.2 PCA on ACWI 286

- 15 Options 287**
 - 15.1 European Options 287
 - 15.2 American Options 301
 - 15.3 Embedded Optionality in Callable Bonds 302
 - 15.4 Exercises 305
 - 15.4.1 Black-Scholes 305
 - 15.4.2 Plot d1 as a Function of Time 306

- 16 Value at Risk 307**
 - 16.1 Parametric VaR 309
 - 16.2 Nonparametric VaR 311
 - 16.3 Calculating VaR Using the Covariance Matrix 312
 - 16.4 Conditional Value at Risk 313
 - 16.5 Calculating VaR Using PerformanceAnalytics 314
 - 16.6 Calculating VaR Using Tidyquant 315
 - 16.7 Chapter-End Summary 318
 - 16.8 Exercises 319
 - 16.8.1 How Sensitive is VaR to α , Revisited 319
 - 16.8.2 Comparing VaR Methods 319
 - 16.8.3 Comparing CVaR Methods 319
 - 16.8.4 Rolling VaR 320
 - 16.8.5 Non-parametric VaR 320

- 17 Time Series Analysis 321**
 - 17.1 ACFs and PACFs 328
 - 17.2 But What Are These Autoregressive (AR) and Moving Average (MA) Models? 335
 - 17.3 Fitting a Model 336
 - 17.4 Forecasting 340
 - 17.5 First Differencing, or Integrated Model? 342
 - 17.6 A Digression: The Intuition of the ACF Values 343

- 18 Machine Learning 345**
 - 18.1 Supervised Algorithms 345
 - 18.2 KNN 347
 - 18.3 Logistic Regression 351
 - 18.4 Decision Tree 355
 - 18.5 Regression Trees (Supervised) 359
 - 18.6 K-Means Clustering 361

18.7 Hierarchical Clustering	366
18.8 Chapter-End Summary	368
18.9 Exercises	370
18.9.1 K-means Clustering on GICS Industries	370
18.9.2 Hierarchical Clustering on P/CF and ROE	370
19 Presenting the Results of Your Analyses	373
19.1 Markdown Documents	373
19.2 Shiny	377
20 Appendix: Main Packages Seen in this Book	381
Index	383