

Data Understanding, Data Analysis, and Data Science Course Notes

Volume 1: Prelude to Data Understanding

Patrick Boily

October 2023

Quadrangle | Idlewyld Analytics and Consulting Services



This work is licensed under a [Creative Commons Attribution – NonCommercial – ShareAlike 4.0 International License](#) [↗](#).

Below is a human-readable summary of (and not a substitute for) the license. Please see [this page](#) [↗](#) for the full legal text.

You are free to:

Share – copy and redistribute the material in any medium or format

Remix – remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution – You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike – If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions – You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

This one goes out to the “Welsh” contingent: Elowyn, Llewellyn, and Gwynneth. Your world is going to be a whole lot different than mine was; maybe data can even help make some of it better. But one thing’s for sure: data is not going away any time soon – better be prepared.

Series Preface

The *first* thing to know about *Data Understanding, Data Analysis, and Data Science* (DUDADS) is that it isn't really a "book". It makes more sense to think of it as **course notes**, or as a **reference manual** and a source of examples and application.

I borrow some of its contents from authors who do a better job of explaining things than I could hope to do; I also sometimes modify their examples and code to better suit my pedagogical needs.* Major influences include [1, 2, 3, 4, 5, 6, 8] – be sure to give these masterful works the attention they deserve!

The *second* thing to know about DUDADS is that it isn't really "a" book. It makes more sense to think of it as **a bunch of books in a trench coat, masquerading as single one**.[†] No one is expected to traverse DUDADS in one sitting, or even to tackle more than a few of its assigned chapters, sections, subsections, exercises at any given time; rather, it is intended to be read in parallel with guided lectures.

The *third* thing to know about DUDADS is that the practical examples use R and/or Python, for no particular reason other than that *some* programming language had to be used to illustrate the concepts. In the text, R code appears in blue boxes:

```
... some R code ...
```

Whereas Python code appears in green boxes:

```
... some Python code ...
```

You may look at some piece of code and think to yourself: "This isn't how I would do it" or "such-and-such a task would be easier to accomplish if we used module/package ABC or programming language XYZ". That's quite possible.

But finding the optimal tool is not the point of this book. In the first place, new data science tools appear regularly, and it would be a fool's errand to try to continuously modify the book to keep up with them.[‡] In the second place, I am serious about the "Understanding" part of *Data Understanding, Data Analysis, and Data Science*, and that is why I favour a **tool-agnostic** approach.

* In all cases, I have attempted to properly cite and give credit where it is due. Get in touch if you find omissions!

[†] I paid heed to this realization by splitting it into a number of volumes.

[‡] I am not saying that I won't be adding examples in different languages in the future, but let's not get ahead of ourselves.

The *fourth* thing to know about DUDADS is that it is not a place to go to in order to obtain a detailed step-by-step guide on “how to solve it”. In person, my answer to a vast array of data science related questions is, rather anti-climatically: “it depends”. Of course, it depends; on the data, on the objectives, on the cost associated with making a mistake, on the stakeholder’s appetite for uncertainty, and, perhaps more surprisingly, on the analytical and data preparation choices that are made along the way.

To some, this might smack of post-modernism: “you are saying that there is no truth, and that data analysis is pointless!” To which I respond: “analysts have agency (lots of it, it turns out), and their choices *DO* influence the results, so make sure to run multiple analyses to determine the variability of the outcomes”. That is the nature of the discipline.

The *last* thing you should probably know about DUDADS is that I have made a concerted effort to focus mainly on the **story** of (learning) data analysis and data science; sometimes, that comes at the expense of rigorous exposition.

“The early stages of education have to include a lot of lies-to-children, because early explanations have to be simple. However, we live in a complex world, and lies-to-children must **eventually be replaced** by more complex stories if they are not to become delayed-action genuine lies.” [7]

Some of the concepts and notions that I present are **incomplete** by design, but remain (I hope) true-to-their-spirit, or at least true “enough” for a first pass.[§] My position is that learning is an iterative process and that important take-aways from an early stage might need to be modified to account for new developments at a later date. But all things in good time: flexibility is a friend in your learning adventure; perfectionism, not always so.

Patrick Boily
Wakefield, October 2023

The DUDADS reference manuals are available at idlewylldanalytics.com [↗](#)

- Volume 1: *Prelude to Data Understanding*
- Volume 2: *Fundamentals of Data Insight*
- Volume 3: *Spotlight on Machine Learning*
- Volume 4: *Techniques of Data Analysis*
- Volume 5: *Special Topics in Data Science and Artificial Intelligence*
- *The Practice of Data Visualization* (with S. Davies and J. Schellinck)

[§] In the parlance of the field, let me simply say that some of the details are left as an exercise for the reader (and can also be found in the numerous references).

Preface References

- [1] C.C. Aggarwal. *Data Mining: the Textbook* [↗](#) . Cham: Springer, 2015.
- [2] C.C. Aggarwal, ed. *Data Classification: Algorithms and Applications* [↗](#) . CRC Press, 2015.
- [3] C.C. Aggarwal and C.K. Reddy, eds. *Data Clustering: Algorithms and Applications* [↗](#) . CRC Press, 2014.
- [4] D. Dalpiaz. *R for Statistical Learning* [↗](#) . 2020.
- [5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* [↗](#) , 2nd ed. Springer, 2008.
- [6] G. James et al. *An Introduction to Statistical Learning: With Applications in R* [↗](#) . Springer, 2014.
- [7] I. Stewart, J. Cohen, and T. Pratchett. *The Science Of Discworld*. Ebury Publishing, 2002.
- [8] H. Wickham and G. Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* [↗](#) . O'Reilly, Jan. 2017.

Contributors and Influences

A reference manual of this size could not have been compiled without the help of a multitude of individuals over the years, who provided contributions, influences, and/or inspiration:

Kevin Cheung *Optimization, Programming Primer*

Aidan Crowther *Programming Primer*

Benoit Dionne *Basics of Numerical Methods*

Fabrizio Donzelli *Overview of Linear Algebra, Multivariate Calculus for Data Analysis*

Patrick Farrell *Survey Sampling Methods*

Ehssan Ghashim *Programming Primer*

Diane Guignard *Basics of Numerical Methods*

Shintaro Hagiwara *Introductory Statistical Analysis*

David Haziza *Design of Experiments*

Rafal Kulik *Probability and Applications, Introductory Statistical Analysis, Classical Regression Analysis, Time Series and Forecasting*

Gilles Lamothe *Classical Regression Analysis*

Dong Elle Liu *Time Series and Forecasting*

Landon Liu *Time Series and Forecasting*

Chunyun Ma *Programming Primer*

Jen Schellinck *Programming Primer, Simulations and Modeling*

A hearty “thank you” to everyone, and to all others with whom I have crossed paths on this data adventure!

Learning Paths

I mostly use the material found in this volume at different levels in my teaching at the University of Ottawa in the Department of Mathematics and Statistics.

In particular, here is what I cover in various courses:

- **MAT 2377** (*Probability and Statistics for Engineers*) – Chapters 6–7, Section 8.2;
- **MAT 3375** (*Regression Analysis*) – Chapter 8 (and some material from Chapter 20, in DUDADS Volume 3);
- **MAT 3377** (*Sampling and Surveys*) – Chapter 10;
- **MAT 3378** (*Analysis of Experimental Design*) – Chapter 11;
- **MAT 3379** (*Introduction to Time Series Analysis*) – Chapter 9.

In 3rd-year courses (and above), I further assume that students are familiar with the contents of Chapters 1–7.