Data Product

May, 2023

Table of Contents

# About this Course

## 0.1 Available course formats

This course is available in multiple formats which allows you to take it in the way that best suites your needs. You can take it for certificate which can be for free or fee.

* The material for this course can be viewed without login requirement on this [Bookdown website](LINK%20HERE). This format might be most appropriate for you if you rely on screen-reader technology.
* This course can be taken for [free certification through Leanpub](LINK%20HERE).
* This course can be taken on [Coursera for certification here](LINK%20HERE) (but it is not available for free on Coursera).
* Our courses are open source, you can find the [source material for this course on GitHub](LINK%20HERE).

# 1 First week

In this overview module, we’ll go over some information and resources to help you get started and succeed in the course.

## 1.1 Welcome

### 1.1.1 Welcome to Developing Data Products

### 1.1.2 A Note of Explanation

This course is used in two different specializations. The Data Science Specialization is a 10-course sequence that covers the full length of the data science pipeline, whereas Data Science: Statistics and Machine Learning covers only the second half of that sequence.

If you are taking Data Science: Statistics and Machine Learning, please note that you will occasionally encounter references to the full 10-course Data Science Specialization. We apologize for any confusion that those references may cause.

If you find that you are not quite ready for the courses in Data Science: Statistics and Machine Learning, you should consider first completing Data Science: Foundations using R in order to build the foundational skills that will help you be prepared for the next step of the journey.

### 1.1.3 Syllabus

Syllabus - [Read on GitHub](https://datasciencespecialization.github.io/Developing_Data_Products/syllabus.html)

#### 1.1.3.1 Course Title

Developing Data Products

#### 1.1.3.2 Course Instructor(s)

The primary instructor of this class is [Brian Caffo](http://www.bcaffo.com/)

Brian is a professor at Johns Hopkins Biostatistics and co-directs the [SMART working group](http://www.smart-stats.org/)

This class is co-taught by Roger Peng and Jeff Leek.

#### 1.1.3.3 Course Description

A data product is the production output from a statistical analysis. Data products automate complex analysis tasks or use technology to expand the utility of a data informed model, algorithm or inference. This course covers the basics of creating data products using Shiny, R packages, and interactive graphics. The course will focus on the fundamentals of creating a data product that can be used to tell a story about data to a mass audience.

In this class students will learn a variety of core tools for creating data products in R and R Studio in specific. Students will be evaluated via quizzes and a culminating project.

Course Content

The lectures will be taught over four weeks with the third week dedicated to creating R packages.

The weeks are organized as follows

1. Shiny, rCharts, manipulate, googleVis
2. Presenting data analysis, slidify, R Studio presenter.
3. Students creating and deploying their projects
4. Creating R packages, classes and methods, yhat.

#### 1.1.3.4 Github repository

The most up to date information on the course lecture notes will always be in the Github repository

The data science specialization is here

<https://github.com/DataScienceSpecialization/Developing_Data_Products>

Please issue pull requests so that we may improve the materials.

#### 1.1.3.5 YouTube

If you’d prefer to watch the videos on YouTube, most of them can be found here:

<https://www.youtube.com/playlist?list=PLpl-gQkQivXhr9PyOWSA3aOHf4ZNTrs90>

#### 1.1.3.6 Book: Developing Data Products in R

This book introduces the topic of Developing Data Products in R. A data product is the ideal output of a Data Science experiment. This book is based on the Coursera Class “Developing Data Products” as part of the Data Science Specialization. Particular emphasis is paid to developing Shiny apps and interactive graphics.

The book is available here: <https://leanpub.com/ddp>

It’s variable pricing, including free! It also includes some content (like leaflet) that was not covered in the class and omits some other. It’s a little rough, but as I work on it you’ll get all of the updates.

#### 1.1.3.7 Weekly quizzes

* There are three weekly quizzes.
* You must earn a grade of at least 80% to pass a quiz
* You may attempt each quiz up to 3 times in 8 hours.
* The score from your most successful attempt will count toward your final grade.

#### 1.1.3.8 Course Project

The Course Project is an opportunity to demonstrate the skills you have learned during the course. It is graded through peer assessment. You must earn a grade of at least 80% to pass the course project.

#### 1.1.3.9 Grading policy

You must score at least 80% on all assignments (Quizzes & Project) to pass the course.

Your final grade will be calculated as follows:

* Quiz 1 = 20%
* Quiz 2 = 20%
* Quiz 3 = 20%
* Course project = 40%

#### 1.1.3.10 Differences of opinion

Keep in mind that currently data analysis is as much art as it is science - so we may have a difference of opinion - and that is ok! Please refrain from angry, sarcastic, or abusive comments on the message boards. Our goal is to create a supportive community that helps the learning of all students, from the most advanced to those who are just seeing this material for the first time.

#### 1.1.3.11 Plagiarism

Johns Hopkins University defines plagiarism as “…taking for one’s own use the words, ideas, concepts or data of another without proper attribution. Plagiarism includes both direct use or paraphrasing of the words, thoughts, or concepts of another without proper attribution.” We take plagiarism very seriously, as does Johns Hopkins University.

We recognize that many students may not have a clear understanding of what plagiarism is or why it is wrong. Please see the JHU referencing guide for more information on plagiarism.

It is critically important that you give people/sources credit when you use their words or ideas. If you do not give proper credit – particularly when quoting directly from a source – you violate the trust of your fellow students.

The Coursera Honor code includes an explicit statement about plagiarism:

I will register for only one account. My answers to homework, quizzes and exams will be my own work (except for assignments that explicitly permit collaboration). I will not make solutions to homework, quizzes or exams available to anyone else. This includes both solutions written by me, as well as any official solutions provided by the course staff. I will not engage in any other activities that will dishonestly improve my results or dishonestly improve/hurt the results of others.

#### 1.1.3.12 Reporting plagiarism on course projects

One of the criteria in the project rubric focuses on plagiarism. Keep in mind that some components of the projects will be very similar across terms and so answers that appear similar may be honest coincidences. However, we would appreciate if you do a basic check for obvious plagiarism and report it during your peer assessment phase.

It is currently very difficult to prove or disprove a charge of plagiarism in the MOOC peer assessment setting. We are not in a position to evaluate whether or not a submission actually constitutes plagiarism, and we will not be able to entertain appeals or to alter any grades that have been assigned through the peer evaluation system.

But if you take the time to report suspected plagiarism, this will help us to understand the extent of the problem and work with Coursera to address critical issues with the current system.

### 1.1.4 Welcome

Welcome - [Read on GitHub](https://datasciencespecialization.github.io/Developing_Data_Products/welcome.html)

I’m glad that you decided to take Developing Data Products, part of the Data Science Specialization from Johns Hopkins Biostatistics!

A data product is the production output from a statistical analysis. Data products automate complex analysis tasks or use technology to expand the utility of a data informed model, algorithm or inference. This course covers the basics of creating data products using Shiny, R packages, and interactive graphics. This course focuses on the statistical fundamentals of creating a data product that can be used to tell a story about data to a mass audience.

You will learn how to communicate using statistics and statistical products. Emphasis will be paid to communicating uncertainty in statistical results. You will learn how to create simple Shiny web applications and R packages for their data products. In addition, we’ll cover reproducible presentations and interactive graphics.

We believe that the key word in Data Science is “science”. Our specialization is focused on providing you with three things: (1) an introduction to the key ideas behind working with data in a scientific way that will produce new and reproducible insight, (2) an introduction to the tools that will allow you to execute on a data analytic strategy, from raw data in a database to a completed report with interactive graphics, and (3) on giving you plenty of hands on practice so you can learn the techniques for yourself. This course represents the final cog in a data science application, creating an end-usable data product.

We are excited about the opportunity to attempt to scale Data Science education. We intend for the courses to be self-contained, fast-paced, and interactive.

Some Basics

A couple of first week housekeeping items. First, make sure that you’ve had R Programming and the Data Scientist’s Toolbox. Reproducible Research would be helpful, but is not mandatory. At a minimum you must know: very basic git, basic R and very basic knitr.

An important aspect of this class is to peruse the materials in the github repository. All of the most up to date material can be found here: <https://github.com/DataScienceSpecialization/Developing_Data_Products>

You should clone this repository as your first step in this class and make sure to fetch updates periodically. (Please send pull requests too!) It is one of the most essential components of the Specialization that you start to use Git frequently. We’re practicing what we preach as well by using the tools in the series to create the series, especially git.

You can clone the whole repo with (http)

git clone <https://github.com/DataScienceSpecialization/Developing_Data_Products.git> or (ssh) git clone [git@github.com](mailto:git@github.com):DataScienceSpecialization/Developing\_Data\_Products.git The lectures are in the index.Rmd lecture files. In this class, we’ll cover how to create these sorts of slides. You will see all of the R code to recreate the lectures. Going through the R code is the best way to familiarize yourself with the lecture materials.

The lecture material for this class is largely front-loaded. This is because the latter time of the class is devoted to developing your data application. Thus the class should be doable in about a 1 month’s time or maybe less. Though make sure you’re keeping up with the classes at the beginning so that you have some space in your schedule later on for app development!

If you’d like to keep up with the instructors I’m ([**bcaffo?**](#ref-bcaffo)) on twitter, Roger is ([**rdpeng?**](#ref-rdpeng)) and Jeff is ([**jtleek?**](#ref-jtleek)). The Department of Biostat here is ([**jhubiostat?**](#ref-jhubiostat)).

<https://datasciencespecialization.github.io/Developing_Data_Products/welcome.html>

### 1.1.5 Book: Developing Data Products in R

This companion book for the class introduces the topic of Developing Data Products in R. You can get a copy here: <https://leanpub.com/ddp>

It has variable pricing, including free!

The book is available under variable pricing, including free! It also includes some content (like leaflet) that was not covered in the class and omits some other. It’s a little rough, but as I work on it you’ll get all of the updates.

### 1.1.6 Community Site - [Read on GitHub](https://datasciencespecialization.github.io/Developing_Data_Products/community.html)

Since the beginning of the Data Science Specialization, we’ve noticed the unbelievable passion students have about our courses and the generosity they show toward each other on the course forums. A couple students have created quality content around the subjects we discuss, and many of these materials are so good we feel that they should be shared with all of our students.

We’re excited to announce that we’ve created a site using GitHub Pages: <http://datasciencespecialization.github.io/> to serve as a directory for content that the community has created. If you’ve created materials relating to any of the courses in the Data Science Specialization, please send us a pull request so we can add a link to your content on our site. You can find out more about contributing here: <https://github.com/DataScienceSpecialization/DataScienceSpecialization.github.io#contributing>

We can’t wait to see what you’ve created and where the community can take this site! ### R and RStudio Links & Tutorials [Practice with Lab Sandbox](https://www.coursera.org/learn/data-products/lab-sandbox?path=%2F)

If you haven’t yet installed R and RStudio, you’ll need to do so now. Here are some links and video tutorials.

Links

[Rstudio](https://posit.co) [R project](https://www.r-project.org)

Tutorials

[Installing R for Windows](https://www.youtube.com/watch?v=LII6of-5Odw) [Installing R for Mac](https://www.youtube.com/watch?v=xokJUwn0mis) [Installing RStudio Mac](https://www.youtube.com/watch?v=JbTMvQ-SbvQ)

## 1.2 Shiny Part 1

### 1.2.1 Shiny - [Read on GitHub](https://datasciencespecialization.github.io/Developing_Data_Products/shiny.html)

**Remember to read the book chapters on Shiny.**

Shiny is an important enough topic to devote a lot of time to it. Shiny is another product by RStudio and it is described by RStudio as “A web application framework for R”. They further add “Turn your analyses into interactive web applications No HTML, CSS, or JavaScript knowledge required”. This is mostly true, though a little HTML at least would be useful for understanding some of the concepts. Here’s a useful site for learning html basics. We’ll proceed as if your html knowledge is very basic and no more advanced than understanding heading levels for fonts. It is important to dinstiguish between a Shiny applications (app) and a Shiny server. A Shiny server is required to host a shiny app for the world. Otherwise, only those who have have shiny installed and have access to your code could run your web page (really defeating the purpose of making a web page in the first place).

In this class, we won’t cover creating a shiny server, as that requires understanding a little linux server administration. Instead, we’ll run our apps locally and use RStudio’s service for hosting shiny apps (their servers) on a platform called shinyapps.io. In other words, RStudio does the server work for your so that all you need to worry about is building your app. Shinyapps.io is free up to a point in that you can only run 5 apps for a certain amount of time per month. This will be fine for our purposes, but if you’re really going to get into making Shiny apps, you’ll have to spring for a paid plan or run your own server.

### 1.2.2 Shinyapps.io Project

**Shinyapps.io Project -** [**Read on GitHub**](https://datasciencespecialization.github.io/Developing_Data_Products/shinyproject.html)

Some people in this session let us know that they are concerned about running up against the 25-hour per month limit on the free tier of shinyapps.io.

Should you hit the limit on the free plan, RStudio will send you a message. If you receive the message and are more than a few days from getting a fresh 25 hours on your monthly renewal, please send an email to [shinyapps-support@rstudio.com](mailto:shinyapps-support@rstudio.com) with the email address you use on the service and the account name you are using (the first part of the URL). RStudio will then increase your limit so you can continue working on your project.

Since there are a lot of folks in the class we’d appreciate if you only emailed RStudio after you get the message and only if you feel you’ll need more time.

### 1.2.3 Shiny 1.1

### 1.2.4 Shiny 1.2

### 1.2.5 Shiny 1.3

### 1.2.6 Shiny 1.4

### 1.2.7 Shiny 1.5

## 1.3 Shiny Part 2

### 1.3.1 Shiny 2.1

### 1.3.2 Shiny 2.2

### 1.3.3 Shiny 2.3

### 1.3.4 Shiny 2.4

### 1.3.5 Shiny 2.5

### 1.3.6 Shiny 2.6

## 1.4 Shiny Gadgets

### 1.4.1 Shiny Gadgets 1.1

### 1.4.2 Shiny Gadgets 1.2

### 1.4.3 Shiny Gadgets 1.3

## 1.5 GoogleVis

### 1.5.1 GoogleVis 1.1

### 1.5.2 GoogleVis 1.2

## 1.6 Plotly

### 1.6.1 Plotly 1.1

### 1.6.2 Plotly 1.2

[](<https://d3c33hcgiwev3.cloudfront.net/3Ob1cF8eEeazDA6HAv4GcQ.processed/full/540p/index.webm?Expires=1683849600&>

### 1.6.3 Plotly 1.3

### 1.6.4 Plotly 1.4

### 1.6.5 Plotly 1.5

### 1.6.6 Plotly 1.6

### 1.6.7 Plotly 1.7

### 1.6.8 Plotly 1.8

## 1.7 Quiz 1

Have to add the quiz here!

# About the Authors

These credits are based on our [course contributors table guidelines](https://www.ottrproject.org/more_features.html#giving-credits-to-contributors).

| Credits | Names |
| --- | --- |
| **Pedagogy** |  |
| Lead Content Instructor(s) | [FirstName LastName](link%20to%20personal%20website) |
| Lecturer(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | Delivered the course in some way - video or audio |
| Content Author(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | If any other authors besides lead instructor |
| Content Contributor(s) (include section name/link in parentheses) - make new line if more than one section involved | Wrote less than a chapter |
| Content Editor(s)/Reviewer(s) | Checked your content |
| Content Director(s) | Helped guide the content direction |
| Content Consultants (include chapter name/link in parentheses or word “General”) - make new line if more than one chapter involved | Gave high level advice on content |
| Acknowledgments | Gave small assistance to content but not to the level of consulting |
| **Production** |  |
| Content Publisher(s) | Helped with publishing platform |
| Content Publishing Reviewer(s) | Reviewed overall content and aesthetics on publishing platform |
| **Technical** |  |
| Course Publishing Engineer(s) | Helped with the code for the technical aspects related to the specific course generation |
| Template Publishing Engineers | [Candace Savonen](https://www.cansavvy.com/), [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/) |
| Publishing Maintenance Engineer | [Candace Savonen](https://www.cansavvy.com/) |
| Technical Publishing Stylists | [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/), [Candace Savonen](https://www.cansavvy.com/) |
| Package Developers ([ottrpal](https://github.com/jhudsl/ottrpal)) [Candace Savonen](https://www.cansavvy.com/), [John Muschelli](https://johnmuschelli.com/), [Carrie Wright](https://carriewright11.github.io/) |  |
| **Art and Design** |  |
| Illustrator(s) | Created graphics for the course |
| Figure Artist(s) | Created figures/plots for course |
| Videographer(s) | Filmed videos |
| Videography Editor(s) | Edited film |
| Audiographer(s) | Recorded audio |
| Audiography Editor(s) | Edited audio recordings |
| **Funding** |  |
| Funder(s) | Institution/individual who funded course including grant number |
| Funding Staff | Staff members who help with funding |

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value   
## version R version 4.0.2 (2020-06-22)  
## os Ubuntu 20.04.5 LTS   
## system x86\_64, linux-gnu   
## ui X11   
## language (EN)   
## collate en\_US.UTF-8   
## ctype en\_US.UTF-8   
## tz Etc/UTC   
## date 2023-05-12   
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date lib source   
## assertthat 0.2.1 2019-03-21 [1] RSPM (R 4.0.5)   
## bookdown 0.24 2023-03-28 [1] Github (rstudio/bookdown@88bc4ea)   
## cachem 1.0.7 2023-02-24 [1] CRAN (R 4.0.2)   
## callr 3.5.0 2020-10-08 [1] RSPM (R 4.0.2)   
## cli 3.6.1 2023-03-23 [1] CRAN (R 4.0.2)   
## crayon 1.3.4 2017-09-16 [1] RSPM (R 4.0.0)   
## desc 1.2.0 2018-05-01 [1] RSPM (R 4.0.3)   
## devtools 2.3.2 2020-09-18 [1] RSPM (R 4.0.3)   
## digest 0.6.25 2020-02-23 [1] RSPM (R 4.0.0)   
## ellipsis 0.3.1 2020-05-15 [1] RSPM (R 4.0.3)   
## evaluate 0.20 2023-01-17 [1] CRAN (R 4.0.2)   
## fastmap 1.1.1 2023-02-24 [1] CRAN (R 4.0.2)   
## fs 1.5.0 2020-07-31 [1] RSPM (R 4.0.3)   
## glue 1.4.2 2020-08-27 [1] RSPM (R 4.0.5)   
## htmltools 0.5.5 2023-03-23 [1] CRAN (R 4.0.2)   
## knitr 1.33 2023-03-28 [1] Github (yihui/knitr@a1052d1)   
## magrittr 2.0.3 2022-03-30 [1] CRAN (R 4.0.2)   
## memoise 2.0.1 2021-11-26 [1] CRAN (R 4.0.2)   
## pkgbuild 1.1.0 2020-07-13 [1] RSPM (R 4.0.2)   
## pkgload 1.1.0 2020-05-29 [1] RSPM (R 4.0.3)   
## prettyunits 1.1.1 2020-01-24 [1] RSPM (R 4.0.3)   
## processx 3.4.4 2020-09-03 [1] RSPM (R 4.0.2)   
## ps 1.4.0 2020-10-07 [1] RSPM (R 4.0.2)   
## R6 2.4.1 2019-11-12 [1] RSPM (R 4.0.0)   
## remotes 2.2.0 2020-07-21 [1] RSPM (R 4.0.3)   
## rlang 1.1.0 2023-03-14 [1] CRAN (R 4.0.2)   
## rmarkdown 2.10 2023-03-28 [1] Github (rstudio/rmarkdown@02d3c25)  
## rprojroot 2.0.3 2022-04-02 [1] CRAN (R 4.0.2)   
## sessioninfo 1.1.1 2018-11-05 [1] RSPM (R 4.0.3)   
## stringi 1.5.3 2020-09-09 [1] RSPM (R 4.0.3)   
## stringr 1.4.0 2019-02-10 [1] RSPM (R 4.0.3)   
## testthat 3.0.1 2023-03-28 [1] Github (R-lib/testthat@e99155a)   
## usethis 1.6.3 2020-09-17 [1] RSPM (R 4.0.2)   
## withr 2.3.0 2020-09-22 [1] RSPM (R 4.0.2)   
## xfun 0.26 2023-03-28 [1] Github (yihui/xfun@74c2a66)   
## yaml 2.2.1 2020-02-01 [1] RSPM (R 4.0.3)   
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library

# 2 References