'm very excited to start R Programming and I hope you are too. This is the second course in the [Data Science Specialization](https://www.coursera.org/specialization/jhudatascience/1) and it focuses on the nuts and bolts of using R as a programming language.

The recommended background for this course is the course [The Data Scientist's Toolbox](https://www.coursera.org/course/datascitoolbox). It is possible to take this class concurrently with that class but you may have to read ahead in the prerequisite class to get the relevant background for this class. For a complete set of course dependencies in the Data Science Specialization please see the [course dependency chart](https://d396qusza40orc.cloudfront.net/rprog/doc/JHDSS_CourseDependencies.pdf).

As you browse the course web site, please make sure to read through the **syllabus** which contains important information about the grading policy.

Please pay particular attention to the differences among the various Programming Assignments. Whereas Programming Assignments 1 and 3 are graded via unit tests that use a submission script that will compare the output of your functions to the correct output, Programming Assignment 2 requires that you submit R code for evaluation and grading by your classmates.

The primary way to interact with me and the other students in this course is through the **discussion forums**. Here, you can start new threads by asking questions or you can respond to other people's questions. If you have a question about any aspect of the course, I strongly suggest that you search through the discussion boards first to see if anyone has already asked that question. If you see something similar to what you want to ask, you should up-vote that question using the up-arrow button rather than asking your question separately. The more votes a question or comment gets, the more likely it is that I will see it and be able to respond quickly. Of course, if you don't see a question similar to the one you want to ask, then you should definitely start a new thread on the appropriate forum.

Finally, consider getting the course textbook, [*R Programming for Data Science*](http://leanpub.com/rprogramming?utm_source=coursera&utm_medium=CourseraEmail&utm_campaign=Coursera), which is available for free from Leanpub. The content in the book tracks the material covered in the course and allows you to hang on to the material once the course is finished.

I hope you enjoy the class. I anticipate a fun four weeks!

Roger Peng and the Data Science Team

**Instructor**

[Roger D. Peng](http://www.biostat.jhsph.edu/~rpeng/) is an Associate Professor of [Biostatistics](http://www.biostat.jhsph.edu/) at the Johns Hopkins Bloomberg School of Public Health and a Co-Editor of the [Simply Statistics blog](http://simplystatistics.org/). He conducts research in the areas of air pollution, climate change, and health risk assessment and statistical methods for spatial and temporal data. He created the course Statistical Programming at Johns Hopkins where it has been taught for the past 8 years. He is also a national leader in the area of methods and standards for reproducible research and is the Reproducible Research editor for the journal *Biostatistics*. His research is highly interdisciplinary and his work has been published in major substantive and statistical journals, including the *Journal of the American Medical Association*, *Journal of the American Statistical Association*, *Journal of the Royal Statistical Society*, and the *American Journal of Epidemiology*. Dr. Peng is the author of more than a dozen software packages implementing statistical methods for environmental studies, methods for reproducible research, and data distribution tools. He has also given workshops, tutorials, and short courses in statistical computing and data analysis. He received his Ph.D. in Statistics from the University of California, Los Angeles and a B.S. in Applied Mathematics from Yale University where he was a member of Calhoun College.

# R Programming

[Roger D. Peng](https://class.coursera.org/rprog-001/wiki/About_the_Instructor)

## Course Description

In this course you will learn how to program in R and how to use R for effective data analysis. You will learn how to install and configure software necessary for a statistical programming environment, discuss generic programming language concepts as they are implemented in a high-level statistical language. The course covers practical issues in statistical computing which includes programming in R, reading data into R, accessing R packages, writing R functions, debugging, and organizing and commenting R code. Topics in statistical data analysis and optimization will provide working examples.

## Course Content

* Week 1: Overview of R, R data types and objects, reading and writing data
* Week 2: Control structures, functions, scoping rules, dates and times
* Week 3: Loop functions, debugging tools
* Week 4: Simulation, code profiling

## Background lectures

Background lectures about the content of the course with respect to other quantitative courses, course logistics, and the R programming language are provided as reference material. It is not necessary to watch the videos to complete the course, however you may find them useful.

### Course Textbook

The book [R Programming for Data Science](http://bit.ly/rprogrammingcoursera) covers all of the lecture material in this course.

## Assessments

### Quizzes

* There are four 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.

### Programming Assignments

* There are **three required** programming assignments.
* You must earn a grade of at least 80% to pass a programming assignment
* Programming Assignments 1 and 3 are graded via unit tests using a submission script that compares the output of your functions to the correct output.
* Programming Assignment 2 is submitted differently and graded via a peer review.

### swirl Programming Assignment (practice)

* In this course, you have the option to use the [swirl R package](http://swirlstats.com/) to practice some of the concepts we cover in lectures.
* While these lessons will give you valuable practice and you are encouraged to complete as many as possible, please note that they are **completely optional** and you can get full marks in the class without completing them.

## Grading Policy

You must score at least 80% on all required assignments (Quizzes & Programming Assignments) to pass the course. Your final grade will be calculated as follows:

* Week 1 Quiz - 20%
* Week 2 Quiz - 10%
* Week 3 Quiz - 5%
* Week 4 Quiz - 10%
* Programming Assignment 1 (Air Pollution) - 20%
* Programming Assignment 2 (Lexical Scoping) - 10%
* Programming Assignment 3 (Hospital Quality) - 25%
* swirl Programming Assignment (practice) - 0%

## Anonymity

As part of this class you will be required to set up a [GitHub account](https://github.com/). Github is a tool for collaborative code sharing and editing. During this course and other courses in the track you will be submitting links to files you publicly place in your Github account as part of peer evaluation. If you are concerned about preserving your anonymity you should set up an anonymous Github account and be careful not to include any information you do not want made available to peer evaluators.

## Typos

* We are prone to a typo or two - please report them and we will try to update the notes accordingly.
* In some cases, the videos may still contain typos that have been fixed in the lecture notes. The lecture notes represent the most up-to-date version of the course material.

## Differences of opinion

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.

## 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 following guide for more information on plagiarism:

[J](http://www.jhsph.edu/academics/degree-programs/master-of-public-health/current-students/JHSPH-ReferencingHandbook.pdf)[HU Student Handbook on Referencing](http://www.jhsph.edu/academics/degree-programs/master-of-public-health/current-students/JHSPH-StudentReferencing_handbook.pdf)

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.

## Reporting plagiarism on course projects

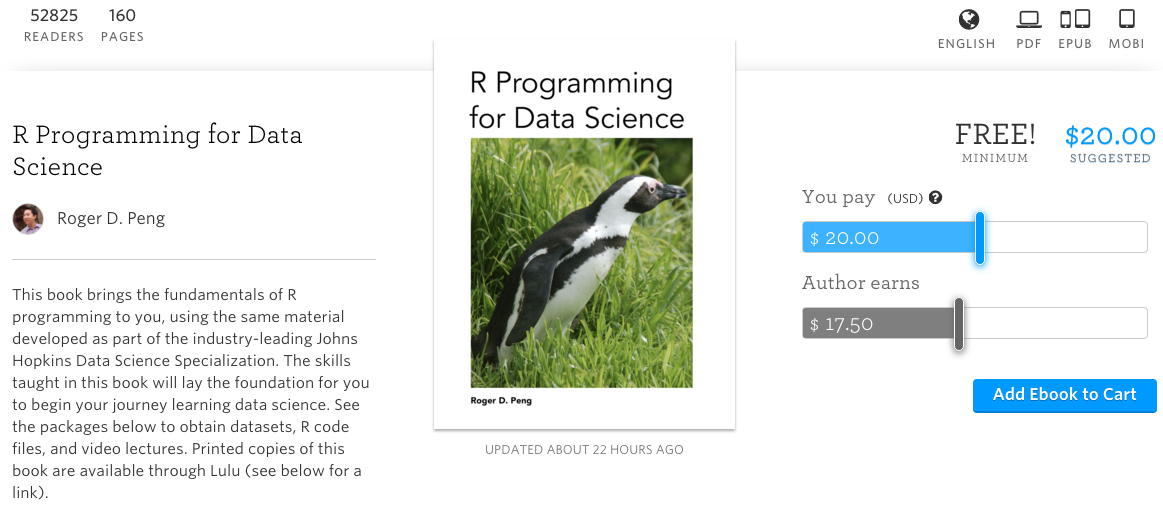
One of the criteria in the project rubric focuses on plagiarism. Keep in mind that some components of the projects will be very similar across terms and so answers that appear similar may be honest coincidences. However, we would appreciate if you do a basic check for obvious plagiarism and report it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

I'm happy to announce that [*R Programming for Data Science*](http://leanpub.com/rprogramming?utm_source=coursera&utm_medium=CourseraEmail&utm_campaign=Coursera) is available through Leanpub. This e-book is specifically designed for use with this course and compiles together all of lecture notes for the course in one place. I think you'll find it useful both for completing this course and for completing the rest of the Data Science Specialization. In addition to allowing you to have all the lecture content in one place, the e-book is a way for you to access the course materials after the class has finished. If you want, printed paper copies of the book are [available from Lulu](http://www.lulu.com/shop/roger-peng/r-programming-for-data-science/paperback/product-22280814.html).

R Programming for Data Science is available in epub (iBooks), mobi (Kindle), and PDF (everything else) formats and so is viewable on just about any computer, tablet, or phone. If you get the book now, you will be entitled to free updates in the future (and there will likely be updates).



In addition to the lecture notes, a compilation of the key lecture videos themselves is available from the Apple iBooks Store via the [*R Programming Course Companion*](http://apple.co/1AAfjiY).

# **Course Supplement: The Art of Data Science**

As you make your way through the world of data science, learning R programming and other important skills, it's important to remember that data science isn't just a collection of tools. It requires a person to apply those tools in a smart way to produce results that are useful to people.

I'm delighted to announce that my latest book, [The Art of Data Science](https://leanpub.com/artofdatascience/), is available from Leanpub (printed paper copies are [available from Lulu](http://www.lulu.com/content/paperback-book/the-art-of-data-science/17220653)). This book addresses the problem of how to think about any data analysis, and how to assemble the tools you've learned and apply them from start to finish. This is the book I wish I'd had when I was first starting out.

I hope you enjoy the book and find it useful. I welcome any feedback as we are continuously working to improve the book.

# **Data Science Podcast: Not So Standard Deviations**

For regular discussions of the latest data science topics, you can go to the [Not So Standard Deviations](https://soundcloud.com/nssd-podcast) podcast that is co-hosted by myself and [Dr. Hilary Parker](https://www.linkedin.com/in/parkerhilary), a Data Scientist at Stitch Fix. The goal of the podcast is to talk about important data science topics and to have a little fun doing it. We also commonly discuss differences between academia and industry as well as the craft of doing data analysis.

If you're an avid podcast listener, you can [subscribe to the podcast through iTunes](https://itunes.apple.com/us/podcast/not-so-standard-deviations/id1040614570) or any popular podcasting app.

# **Getting Started and R Nuts and Bolts**

This week is all about getting started with R and learning some of the basic details of the language. If you haven't already installed R, you should go to the [R web site](http://www.r-project.org/" \t "_blank) and download R for your platform (Windows, Mac, or Unix/Linux). Also, if you want, you can download [RStudio](http://www.rstudio.com/" \t "_blank), which is a free interactive development environment designed for R that is very useful and we use quite a bit in the Data Science Specialization. I've made some videos to help you along with the installation process:

* Installing R on Windows
* Installing R on a Mac
* Installing R on RStudio (on a Mac)

Before you start using R, one key concept is the **working directory**. This is the directory/folder on your computer where you will store project files, data, and code. It's important that you tell R where the working directory is that you will be using so that it knows where to find the appropriate file (the working directory can be any directory on your computer). These videos tell you how to set your working directory:

* Setting your working directory (Windows)
* Setting your working directory (Mac)

## Learning Objectives

By the end of week 1 you should be able to:

* Install the R and RStudio software packages
* Download and install the swirl package for R
* Describe the history of the S and R programming lectures
* Describe the differences between atomic data types
* Execute basic arithmetic operations
* Subset R objects using the "[", "[[", and "$" operators and logical vectors
* Describe the explicit coercion feature of R
* Remove missing (NA) values from a vector

## Assessments

* Quiz 1 - 80% or better required to pass
* There is **no official graded programming assignment for this week**. However, we have developed a series of practice exercises to get you started with R. These exercises are implemented using the swirl package for R. **The swirl programming assignment is NOT required**.