PH125.2x: Data Science: Visualization - Course Syllabus

Course Instructor

Rafael Irizarry

Course Description

In this second course of nine in the HarvardX Data Science Professional Certificate, we learn the basics of data visualization and exploratory data analysis.

The growing availability of informative datasets and software tools has led to increased reliance on data visualizations across many industries, academia, and government. Data visualization provides a powerful way to communicate data-driven findings, motivate analyses, or detect flaws.

In this course, you will learn the basics of data visualization and exploratory data analysis. We will use three motivating examples and ggplot2, a data visualization package for the statistical programming language R, to code. To learn the very basics, we will start with a somewhat artificial example: heights reported by students. Then we will use two case studies related to world health and economics and another in infectious disease trends in the United States.

It is also important to note that mistakes, biases, systematic errors, and other unexpected problems often lead to data that should be handled with care. The fact that it can be difficult or impossible to notice an error just from the reported results makes data visualization particularly important. This course will explore how failure to discover these problems often leads to flawed analyses and false

discoveries.

HarvardX has partnered with DataCamp for some assignments in R that allow students to program directly in a browser-based interface. You will not need to download any special software to use DataCamp, but an up-to-date browser is recommended. In addition, Verified learners have access to several exercises on the edX platform that require writing code on a local installation of R and entering answers on edX. You can read more about installing R below.<

What you'll learn:

- data visualization principles to better communicate data-driven findings
- how to use ggplot2 to create custom plots
- the weaknesses of several widely used plots and why you should avoid them

New to EdX?

Are you new to edX? Check out edx's Demo Course!

Need help? Visit edX Support via the Support tab or visit the Help Center.

Course Structure

When you join the course, we encourage you to meet your peers, learn the DataCamp platform, and tell us about yourselves and what you hope to get out of the course! You can progress through the material at your own pace.

For Verified learners, The eleven DataCamp programming exercises are worth **90%** of your grade. They show up as 10 total grades. The comprehensive assessments at the end of the course combine to be worth **10%** of your grade.

All other components of the course, such as the the discussion boards, are not for credit.

Certification

In order to receive a Verified Certificate, you must sign up and pay for a Verified

Certificate by the deadline on the course page and earn a passing grade of at least **70%**.

Installing R

To install R, you can download it freely from the Comprehensive R Archive Network (CRAN). but if you need further help you can check out chapter 1 of the textbook.

Research

HarvardX pursues the science of learning. When you participate in this course, you will also participate in research about learning. Read our research statement to learn more.

COURSE OUTLINE

Section 1: Introduction to Data Visualization and Distributions

You will get started with data visualization and distributions in R.

Section 2: Introduction to ggplot2

You will learn how to use ggplot2 to create plots.

Section 3: Summarizing with dplyr

You will learn how to summarize data using dplyr.

Section 4: Gapminder

You will see examples of ggplot2 and dplyr in action with the Gapminder dataset.

Section 5: Data Visualization Principles

You will learn general principles to guide you in developing effective data visualizations.

FAQS - ABOUT THE COURSE

Does this course have any prerequisites?

Yes, this is the second course in the HarvardX Data Science Professional Certificate Series. We strongly recommend that you take the first course in the series before taking this course: Data Science: R Basics.

Do I have to take the courses in sequence?

The courses in the HarvardX Data Science Professional Certificate are designed to be taken in the following order:

1. R Basics
2. Visualization
3. Probability
4. Inference and Modeling
5. Productivity Tools
6. Wrangling
7. Linear Regression
8. Machine Learning
9. Capstone

Each subsequent course assumes familiarity with the content in the preceding courses. Depending on your experience with data science generally and R specifically, you may be able to take the courses out of sequence if you choose.

What is this R language we will be using?

R is a programming language and environment that is used in many fields for statistical analysis. R is also completely free and open source.

Do I need to have a background in statistics or R to take the course?

This is the **second** course in the series. We assume you have either taken the first course or have at least some background in R. The statistics and programming aspects of the series increase in difficulty the farther you progress.

I do not have a background in programming. Can I still take the course?

Yes, we do not assume a background in programming beyond that taught in the first course. You will learn programming skills by completing the exercises. That said, unless you have some familiarity with R, we highly recommend starting with the first course, PH125.1x R Basics.

Is this class challenging?

These courses are taught at the college level. Some of the material, depending on your exposure, may be fairly challenging. However, the first few courses are meant to be a "gentle" introduction to statistics and R. Furthermore, there is also substantial help from the community including a lively discussion board.

Is there a textbook?

Yes, there is a free PDF textbook available here. (Note: The book is "free" in that you can slide the "YOU PAY" scale to $0. You are welcome to pay what you can afford, and there is no advantage in the course to anyone that "purchases" the book for more money.)

There is also an HTML version of the textbook here.

How long does this course take?

That is up to you! It is 5 sections of content for all learners, plus an additional comprehensive assessment for verified learners. Just be aware that you must complete the course by the deadline listed on your course homepage.

I am doing well on the assessments, but when I look under "Progress" I have a very low grade...why?

The grade is calculated based on all of the assessments you have completed ***and*** the assessments that you have not completed (edX says you have a "zero" on those assessments **until** you have attempted them). You will see your overall grade move up as you progress through the course.

FAQS - CERTIFICATES

What is the deadline to sign up for a Verified Certificate?

The deadline is listed on the right side of the course landing page.

How do I earn a certificate?

To earn a certificate, you must sign up for a Verified Certificate by the deadline and earn a grade of at least 70%. When you achieve this score, a view your certificate button will appear on your dashboard. For more information, read the instructions from edX.

How do I upgrade to a verified certificate?

Go to your edX Dashboard (by clicking the edX icon at the top left of this page). Under this course, click the "Challenge Yourself!" link.

FAQS - SOFTWARE, ETC.

How do I get started?

First, you will need to install R onto your machine. You can do that from CRAN.

The latest version is 3.6.0. If you are using an older version of R, it will likely be fine so long as you are using **version 3.2.0 or later**. This course makes use of packages such as "dplyr" which only work if you are running version 3.1.2 or more recent. Depending on your machine you may have to resolve various dependencies, but in most instances it should be straightforward to install R.

What do I do after installing R?

You can begin learning or optionally install RStudio.

RStudio is a graphical user interface for R. RStudio is NOT part of the R language nor is it required in order to complete the course. However, it does provide a nice interface (Professor Irizarry uses RStudio in the videos). Please note that RStudio comes in various commercial flavors. However, it does include a free Open Source Edition. If you are unable to install RStudio for whatever reason, we suggest you skip this step and just continue with the course so long as R is successfully installed on your computer.

How do I install packages?

We will be using and downloading various packages throughout this course and the subsequent courses. However, installing packages is straightforward. For example, if you want to install dslabs you just enter

```
install.packages("dslabs")
```

Please note that the **most common error** when trying to install a package is spelling. For example, if you had typed `"dslab"` instead of `"dslabs"` you would get an error. The **second most common error** is forgetting the quotes.

I installed the package but it is not working.

After installing a package, you must load it. For example, after installing dslabs, you load it by typing:

```
library("dslabs")
```

After you hit enter, if you do not get an error, then you are good to go. Note that after running some packages, you may get a message, but that does not imply there is an error. Also note that when starting a new R session you typically will need to load the package again.

Can I install packages using RStudio?

Yes, you can install most if not all packages via RStudio. Click on the right of the screen on Tools > Install Packages. Then you can enter the name of the package you want. If the package is found, you should be able to directly download it.

I'm trying to install other packages such as rafalib and I get an error (similar to): 'lib = "C:/Program Files/R/R-3.2.2/library"' is not writable

This problem occurs when you don't have administrative privileges to overwrite the file location. We strongly recommend that you fix this by having administrator privileges on the machine you are using. If you are using a public or shared machine, you may have difficulty installing the necessary packages.

I'm having problems installing devtools and other packages. Can you help?

devtools can be hard to install which is why we are not using it in any of the exercises. You do not need it for this course. If you wish to install packages which are not used in the course, we may not be able to help you. We certainly do not want to discourage you from exploring, since there are numerous fabulous packages. However, we are going to focus on material necessary for this course.

When I try to open a file I am getting an error such as cannot open file 'femaleMiceWeights.csv': No such file or directory

Probably because the file is not in your working directory. The file needs to be in the folder or you need to change the working directory to the one containing the file.

I am getting some strange error message in R. Ideas?

If you can not figure out what is wrong, a good idea especially if you have been using R for a long time is to exit and restart. If you are still seeing it after exiting and restarting, let us know (see below).

FAQS - GETTING HELP

My code is not working! Can I ask you about it?

Yes certainly! You need to post to the course discussion board. Please READ BELOW on how to post.

How do I post?

Okay, this is important! READ THIS... really, read it, we explain how to ask questions.

When you ask a question please make sure to:

1. List it as a question.
2. Choose the appropriate Topic Area.
3. Have a clear, specific title.

For example if you have a question on the first exercise in section 1, then state so (e.g., "help with question 1a in section 1") in your question.

Also, you can show us your code (or certainly parts of it, we ask you not to ruin the course for others by carelessly posting solutions). Make sure your code is legible. Even simple code can be difficult and annoying to read if garbled. For example do NOT present your code as such:

sum <- 100 for(i in 1:50) sum <- sum + i sum

Instead please make it neat and format it as code:

```
sum <- 100

for(i in 1:50)

sum <- sum + i

sum
```

To do this, you can insert your code, then highlight it and press Ctrl+K and it should be nice and legible (if not, please fix it using the guidance in the pinned post in the “general” discussion forum).

This not only helps us, the staff, identify your problem, it also helps other students who may have similar questions.

For more information about how to use the discussion forum, check out the edX documentation.

FAQs - PROFESSIONAL CERTIFICATE

How often will the courses be offered?

Courses in the program are offered frequently - so if now isn’t a good time for you to start one of the courses you need as a prerequisite or if you missed a

deadline, there will be another offering of the course you need coming soon!

Does the order of courses in the Professional Certificate Program matter?

Yes, order does matter, particularly for the first four courses in the sequence. For the later courses, depending on your previous experience, you may be able to swap the sequence of some of the courses. The courses are designed to be taken in the following order:

1. R Basics
2. Visualization
3. Probability
4. Inference and Modeling
5. Productivity Tools
6. Wrangling
7. Linear Regression
8. Machine Learning
9. Capstone

Do I need to register for all of the courses at once in order to be eligible for the Professional Certificate?

No! You can take courses individually - once you have obtained an ID Verified Certificate in each course, you will be eligible for the Professional Certificate. If you choose to pre-pay for the entire program, you receive a discount on the total registration cost.