**Pearson** | PROFESSIONAL PROGRAMS

# Professional Certificate in Data Science

edX | **HARVARD** UNIVERSITY

**ONLINE INTERACTIVE COURSE**
₹ 40,000 + Applicable Taxes | **Duration: 6-8 months**

# Overview

The demand for skilled data science practitioners in industry, academia, and government is rapidly growing. The HarvardX Data Science program prepares you with the necessary knowledge base and useful skills to tackle real-world data analysis challenges. The program covers concepts such as probability, inference, regression, and machine learning.

Throughout the program, we will be using the R software environment. You will learn R, statistical concepts and data analysis techniques simultaneously. We believe that you can better retain R knowledge when you learn how to solve a specific problem. Furthermore, HarvardX has partnered with DataCamp for all assignments, which use code checking technology that will permit you to get hands-on practice during the courses.

## What You Will Learn

Fundamental R programming skills

Statistical concepts such as probability, inference, and modelling and how to apply them in practice

Gain experience with the tidyverse, including data visualisation with ggplot2 and data wrangling with dplyr

Become familiar with essential tools for practicing data scientists such as Unix/Linux, git and GitHub, and RStudio

Implement machine learning algorithms

In-depth knowledge of fundamental data science concepts through motivating real-world case studies

In each course, we use motivating case studies, ask specific questions, and learn by answering these through data analysis.

**Case studies include:** Trends in World Health and Economics, US Crime Rates, The Financial Crisis of 2007-2008, Election Forecasting, Building a Baseball Team (inspired by Moneyball), and Movie Recommendation Systems.

## Program Benefits

Jumpstart your career with a Professional Certificate in Data Science

### Cutting Edge Curriculum:

In-depth learning of fundamental concepts through motivating real-world case studies

### For The Industry. By The Industry:

Learn straight from key industry leaders for highest degree of practical relevance

### Career Support:

Assisted resume writing with AI-powered resume reviewing and a Job Board for relevant positions

### Online interactive sessions:

Online hand-holding and Q&A sessions by an IIT alumnus will be held twice during the program

## Job Outlook

- R is listed as a required skill in 64% of data science job postings and was Glassdoor's Best Job in America in 2016 and 2017. (Source: Glassdoor)
- Companies are leveraging the power of data analysis to drive innovation. Google data analysts use R to track trends in ad pricing and illuminate patterns in search data. Pfizer created customised packages for R, so scientists can manipulate their own data.
- 32% of full-time data scientists started learning machine learning or data science through a MOOC, while 27% were self-taught. (Source: Kaggle, 2017)
- Data scientists are few in number and high in demand. (source: TechRepublic)

## Program Vitals

**Expert Instruction:**
9 graduate level courses

**Online Interactive Course**
Two Online Interactive Sessions

**Program Fee:**
₹ 40,000 + Applicable taxes

**Payment by Debit/Credit Card**
Loan facility available

**Course Start Date & Duration:**
30th January, 2021, 6-8 months

**Time Commitment:**
1-2 hours per week

## Meet Your Instructor

**Rafael Irizarry**
Professor of Biostatistics,
Harvard University

# Courses:

## Data Sciences: R Basics

Build a foundation in R and learn how to wrangle, analyse, and visualise data.

**Overview:**
This course will introduce you to the basics of R programming. You can better retain R when you learn it to solve a specific problem, so you'll use a real-world dataset about crime in the United States. You will learn the R skills needed to answer essential questions about differences in crime across the different states.

We'll cover R's functions and data types, then tackle how to operate on vectors and when to use advanced functions like sorting. You'll learn how to apply general programming features like "if-else," and "for loop" commands, and how to wrangle, analyse and visualise data.

Rather than covering every R skill you might need, you'll build a strong foundation to prepare you for the more in-depth courses later in the series, where we cover concepts like probability, inference, regression, and machine learning. We help you develop a skill set that includes R programming, data wrangling with dplyr, data visualisation with  ggplot2, file organisation with UNIX/Linux, version control with git and GitHub, and reproducible document preparation with RStudio. The demand for skilled data science practitioners is rapidly growing, and this series prepares you to tackle real-world data analysis challenges.

## What You Will Learn

Basic R syntax

Foundational R programming concepts such as data types, vectors arithmetic, and indexing

How to perform operations in R including sorting, data wrangling using dplyr, and making plots

## Data Sciences: Visualisation

Learn basic data visualisation principles and how to apply them using ggplot2.

**Overview:**

This course covers the basics of data visualisation and exploratory data analysis. We will use three motivating examples and ggplot2, a data visualisation package for the statistical programming language R. We will start with simple datasets and then graduate to case studies about world health, economics, and infectious disease trends in the United States.

We'll also be looking at how 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 a mistake within a dataset makes data visualisation particularly important.

The growing availability of informative datasets and software tools has led to increased reliance on data visualisations across many areas. Data visualisation provides a powerful way to communicate data-driven findings, motivate analyses, and detect flaws. This course will give you the skills you need to leverage data to reveal valuable insights and advance your career.

### What You Will Learn

Data visualisation principles

How to 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

# Data Sciences: Probability

Learn probability theory - essential for a data scientist - using a case study on the financial crisis of 2007–2008.

**Overview:**

In this course you will learn valuable concepts in probability theory. The motivation for this course is the circumstances surrounding the financial crisis of 2007–2008. Part of what caused this financial crisis was that the risk of some securities sold by financial institutions was underestimated. To begin to understand this very complicated event, we need to understand the basics of probability.

We will introduce important concepts such as random variables, independence, Monte Carlo simulations, expected values, standard errors, and the Central Limit Theorem. These statistical concepts are fundamental to conducting statistical tests on data and understanding whether the data you are analysing is likely occurring due to an experimental method or to chance.

Probability theory is the mathematical foundation of statistical inference which is indispensable for analysing data affected by chance, and thus essential for data scientists.

## What You Will Learn

Important concepts in probability theory including random variables and independence

How to perform a Monte Carlo simulation

The meaning of expected values and standard errors and how to compute them in R

The importance of the Central Limit Theorem

# Data Sciences: Inference & Modelling

Learn inference and modelling, two of the most widely used statistical tools in data analysis.

**Overview:**
Statistical inference and modelling are indispensable for analysing data affected by chance, and thus essential for data scientists. In this course, you will learn these key concepts through a motivating case study on election forecasting.

This course will show you how inference and modelling can be applied to develop the statistical approaches that make polls an effective tool and we'll show you how to do this using R. You will learn concepts necessary to define estimates and margins of errors and learn how you can use these to make predictions relatively well and also provide an estimate of the precision of your forecast.

Once you learn this you will be able to understand two concepts that are ubiquitous in data science: confidence intervals, and p-values. Then, to understand statements about the probability of a candidate winning, you will learn about Bayesian modelling. Finally, at the end of the course, we will put it all together to recreate a simplified version of an election forecast model and apply it to the 2016 election.

## What You Will Learn

The concepts necessary to define estimates and margins of errors of populations, parameters, estimates and standard errors in order to make predictions about data

How to use models to aggregate data from different sources

The very basics of Bayesian statistics and predictive modelling

# Data Sciences: Productivity Tools

Keep your projects organised and produce reproducible reports using GitHub, git, Unix/Linux, and RStudio.

**Overview:**
A typical data analysis project may involve several parts, each including several data files and different scripts with code. Keeping all this organised can be challenging.

This course explains how to use Unix/Linux as a tool for managing files and directories on your computer and how to keep the file system organised. You will be introduced to the version control systems git, a powerful tool for keeping track of changes in your scripts and reports. We also introduce you to GitHub and demonstrate how you can use this service to keep your work in a repository that facilitates collaborations.

Finally, you will learn to write reports in R markdown which permits you to incorporate text and code into a document. We'll put it all together using the powerful integrated desktop environment RStudio.

## Data Sciences: Wrangling

Learn to process and convert raw data into the formats needed for analysis.

**Overview:**
In this course we cover several standard steps of the data wrangling process like importing data into R, tidying data, string processing, HTML parsing, working with dates and times, and text mining. Rarely are all these wrangling steps necessary in a single analysis, but a data scientist will likely face them all at some point.

Very rarely is data easily accessible in a data science project. It's more likely for the data to be in a file, a database, or extracted from documents such as web pages, tweets, or PDFs. In these cases, the first step is to import the data into R and tidy the data, using the tidyverse package. The process that converts data from its raw form to the tidy form is called data wrangling.

This process is a critical step for any data scientist. Knowing how to wrangle and clean data will enable you to make critical insights that would otherwise be hidden.

### What You Will Learn

Importing data into R from different file formats

Web scraping

How to tidy data using the tidyverse package to better facilitate analysis

String processing with regular expressions (regex)

Wrangling data using dplyr

How to work with dates and times as file formats

Text mining

## Data Sciences: Linear Regression

Learn how to use R to implement linear regression, one of the most common statistical modelling approaches in data science.

**Overview:**
Linear regression is commonly used to quantify the relationship between two or more variables. It is also used to adjust for confounding. This course covers how to implement linear regression and adjust for confounding in practice using R.

In data science applications, it is very common to be interested in the relationship between two or more variables. The motivating case study we examine in this course relates to the data-driven approach used to construct baseball teams described in Moneyball. We will try to determine which measured outcomes best predict baseball runs by using linear regression.

We will also examine confounding, where extraneous variables affect the relationship between two or more other variables, leading to spurious associations. Linear regression is a powerful technique for removing confounders, but it is not a magical process. It is essential to understand when it is appropriate to use, and this course will teach you when to apply this technique.

# Data Sciences: Machine Learning

Build a movie recommendation system and learn the science behind one of the most popular and successful data science techniques.

**Overview:**
Perhaps the most popular data science methodologies come from machine learning. What distinguishes machine learning from other computer guided decision processes is that it builds prediction algorithms using data. Some of the most popular products that use machine learning include handwriting readers implemented by the postal service, speech recognition, movie recommendation systems, and spam detectors.

In this course, you will learn popular machine learning algorithms, principal component analysis, and regularisation by building a movie recommendation system.

You will learn about training data, and how to use a set of data to discover potentially predictive relationships. As you build the movie recommendation system, you will learn how to train algorithms using training data so you can predict the outcome for future datasets. You will also learn about overtraining and techniques to avoid it, such as cross-validation. All of these skills are fundamental to machine learning.

# Data Sciences: Capstone

Show what you've learned in the Professional Certificate Program in Data Science.

**Overview:**

To become an expert data scientist you need practice and experience. By completing this capstone project you will get an opportunity to apply the knowledge and skills in R data analysis that you have gained throughout the series. This final project will test your skills in data visualisation, probability, inference and modelling, data wrangling, data organisation, regression, and machine learning.

In this course, you will receive much less guidance from instructors. When you complete the project you will have a data product to show off to potential employers or educational programs, a strong indicator of your expertise in the field of data science.

## What You Will Learn

How to apply the knowledge base and skills learned throughout the series to a real-world problem

How to independently work on a data analysis project

# Value Adds:

**Data Science Fundamentals Course** | 20 Hours of Video Instruction

Learn IT Data Science Fundamentals Course teaches you the foundational concepts, theory, and techniques you need to know to become an effective data scientist. The videos present you with applied, example-driven lessons in Python and its associated ecosystem of libraries, where you get your hands dirty with real datasets and see real results.

This course focuses on the fundamentals of acquiring, parsing, validating, and wrangling data with Python and its associated ecosystem of libraries. After an introduction to Data Science as a field and a primer on the Python programming language, you walk through the data science process by building a simple recommendation system. After this introduction, you dive deeper into each of the specific steps involved in the first half of the data science process–mainly how to acquire, transform, and store data (often referred to as an ETL pipeline). You learn how to download data that is openly accessible on the internet by working with APIs and websites, and how to parse this XML and JSON data. With this structured data, you learn how to build data models, store and query data, and work with relational databases. Along the way, you learn the fundamentals of programing with Python (including object-oriented programming and the standard library) as well as the best practices of building sustainable data science applications.

## Instructor

**Jonathan Dinu**
Ph.D. Candidate
**Researcher & Author**

Jonathan Dinu is an author, researcher, and most importantly, an educator. He is currently pursuing a Ph.D. in Computer Science at Carnegie Mellon's Human Computer Interaction Institute (HCII), where he is working to democratise machine learning and artificial intelligence through interpretable and interactive algorithms. Previously, he founded Zipfian Academy (an immersive data science training program acquired by Galvanize), has taught classes at the University of San Francisco, and has built a Data Visualisation MOOC with Udacity.

## Online Interactive Sessions

Online hand holding and Q&A sessions by an IIT alumnus will be held twice during Harvard's Data Science Program course modules to reinforce understanding and provide ample time for queries.

## Instructor

**Ajay Sood**
Guest Faculty
**BITS Pilani**

Ajay Sood is a Guest Faculty at BITS Pilani and consulting with analytics startups. He has also worked for JP Morgan and IBM. He is credited with 14 patents and 6 publications to his name. Ajay completed his ME from IISc, PGDBA from Symbiosis and BTech from IIT Kanpur.
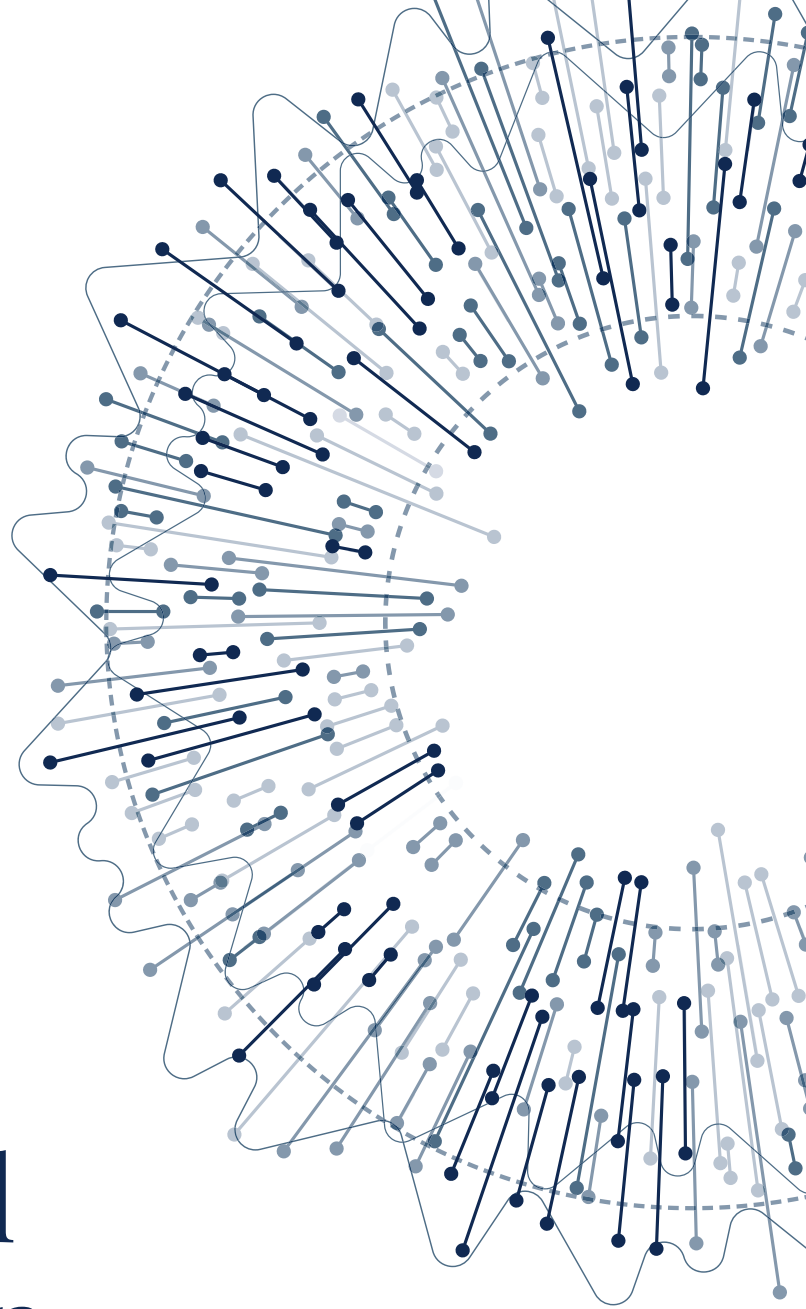
# Professional Certificate in Data Science

## Contact Us

ppp.edge@pearson.com

https://in.pearson.com/ppp

0124 - 4185100

Pearson Professional Programs,
Vatika First Indian Place, Tower C,
Ground Floor (Near M.G. Road
Metro Station), Gurugram, India.