# Data Science – Capture and Explore Data Using R
# Syllabus

2-Days, 9AM - 5PM
Instructor: Dr. Chuck Cartledge, Bojan Duric

Learn and use R to accelerate your Data Scientist career path or to become more efficient and effective in your current role.

During this two-day bootcamp, you will receive a comprehensive hands-on introduction to one of the main tools and ideas in the data scientist's toolbox, the R programming language. You will focus on using R for data acquisition and manipulation to become more efficient in a data science/business analyst setting.

- Objectives:
  - Apply fundamental tidy data concepts
  - Data wrangle R techniques
  - Pull data from csv or web (using API)
  - Perform operations in R including sorting, data wrangling using dplyr, and making plots
  - Apply helper libraries for dates and strings
  - Develop R notebook (reproducible report)
- Technologies to be used:
  - R, and RStudio
- Academic prerequisites:
  - Exposure to, and experience working with excel functions and pivot tables
  - CS-121G or CS150, at least a C as final grade
  - Permission of the instructor
  - Mandatory pre-read material
- Recommended experiences:
  - Exposure to, and experience with R/python/
  - A structured language
- Other notes:
  - This is a hands-on programming course.
  - You will use R and RStudio.
  - You will write simple programs in R
  - Text: R for Data Science, by Garrett Grolemund and Hadley Wickham (ISBN: TBD)

# Table of Contents



## 1 Course Description

The field commonly known as “Data Science” lies at the intersection of mathematics, computer science, and domain expertise. Within the data science (DS) world, there are a multitude of areas of study, and exploration. 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 from City of Virginia Beach Open Data portal. You will learn the R skills needed to answer essential questions and perform basic exploratory analysis.

We will cover R's functions and data types, then tackle how to operate on data frames and when to use data sub-setting techniques. You will learn how to apply general data sub-setting features like “select” and “filter”, and how to wrangle, analyze and visualize data.

Rather than covering every R skill you might need, you will build a strong foundation to prepare you for the more in-depth advanced courses later, concepts like probability, inference, regression, and machine learning. We help you develop a skill set that includes R programming, data wrangling with dplyr, data visualization with ggplot2, and reproducible document preparation with RStudio.

The demand for skilled data science practitioners is rapidly growing, and this course prepares you to tackle real-world data analysis challenges. We will focus primarily on fundamental and most used data wrangling techniques in R language.

## 2 Course Outline

After the completion of this course, students will be able to pull data from different sources (small dataset and large datasets), clean and manipulate data, use rich visualization libraries to deliver your findings as reports and notebooks, and everything from one Open Source platform.

There are three main parts to data wrangling which we will cover:

1. Import
2. Tidy
3. Transform

The bootcamp will include the following:

- Installing and updating R libraries
- Navigating RStudio Integrated Development Environment (IDE)
- Understanding different data types working with R
- Reading/storing data from/in different file types
- Applying "tidyverse" tools in data processing
- Transforming data using dplyr functions (select, filter, group by, summarize, mutate)
- Transforming and manipulating strings with stringr package
- Transforming and using different date formats in analysis using lubridate functions
- Applying grammar of graphics with ggplot2
- Creating reproducible analysis as notebooks and reports (html and/or pdf) in Rmarkdown

## 3 Assignments

There will be individual programming assignments addressing different aspects of R data manipulation as used in data and analytics fields. These include:

1. Satisfactory attendance and participation in the Bootcamp, and
2. Demonstrated R coding technique improvements.

## 4 Grading

### 4.1 Overall grading scale

Overall grade for the course will be based on the student's performance in: class attendance and participation (50%), assignments (50%).

The grading scale follows:

Table 1: Grading scale

| Range Grade | Grade | points |
|---|---|---|
| 94 - 100 | A | 4 |
| 90 - 93 | A- | 3.7 |
| 87 - 89 | B+ | 3.3 |
| 82 - 86 | B | 3 |
| 80 - 81 | B- | 2.7 |
| 77 - 79 | C+ | 2.3 |
| 73 - 76 | C | 2 |
| 70 - 72 | C- | 1.7 |
| 67 - 69 | D+ | 1.3 |
| 63 - 66 | D | 1 |
| 60 - 62 | D- | 0.7 |
| 0 - 59 | F | 0 |
| NA | WF | 0 |

### 4.2 Late Assignments

Assignments are due by midnight of the due date. The time of submission is the timestamp of the e-mail saying that the submission is ready. Assignments that are late will be penalized at the rate of one half of a letter grade per 24-hour period. Late submissions will be accepted up to 4 days late (see Table 2).

Table 2: Late submission maximum grade.

| Hours Late | Max. Grade |
|---|---|
| 0 | A |
| 24 | A- |
| 48 | B+ |
| 72 | B |
| 96 | B- |
| >96 | F |

## 5 Course Policies

### 5.1 Attendance Policy

You are responsible for the contents of all lectures. If you know that you are going to miss a lecture, have a reliable friend take notes for you. Of course, there is no excuse for missing due dates or exam days. During lectures, we will be covering selected material from the textbook. Lectures will consist of

the exploration of real world problems not covered in the book. You will be given a work assignment at the end of each lecture for the next class. Repetition, repetition.

I expect you to attend class and to arrive on time. Your grade may be affected if you are consistently tardy. If you have to miss a class, you are responsible checking the course website to find any assignments or notes you may have missed.

### 5.2 Classroom Conduct

Be respectful of your classmates and instructor by minimizing distractions during class. Cell phones must be turned off during class.

### 5.3 Seeking Help

The course website should be your first reference for questions about the class. Announcements will be posted to the course website. The best way to get help is to research on your own and email with specific challenge to set up an appointment for a Skype conference.

I will be establishing virtual office hours using Skype, and will use Google calendar to coordinate. I am available via email, but do not expect or rely on an immediate response.

### 5.4 Disability Services

In compliance with PL94-142 and more recent federal legislation affirming the rights of disabled individuals, provisions will be made for students with special needs on an individual basis. The student must have been identified as special needs by the university and an appropriate letter must be provided to the course instructor. Provision will be made based upon written guidelines from the University's Office of Educational Accessibility. All students are expected to fulfill all course requirements.

## 6 Academic Integrity / Honor Code

By attending Old Dominion University you have accepted the responsibility to abide by the honor code. If you are uncertain about how the honor code applies to any course activity, you should request clarification from the instructor. The honor pledge is as follows:

"I pledge to support the honor system of Old Dominion University. I will refrain from any form of academic dishonesty or deception, such as cheating or plagiarism. I am aware that as a member of the academic community, it is my responsibility to turn in all suspected violators of the honor system. I will report to Honor Council hearings if I am summoned."

In particular, submitting anything that is not your own work without proper attribution (giving credit to the original author) is plagiarism and is considered to be an honor code violation. It is not acceptable to copy source code or written work from any other source (including other students), unless explicitly allowed in the assignment statement. In cases where using resources such as the Internet is allowed, proper attribution must be given. Any evidence of an honor code violation (cheating) will result in a 0 grade for the assignment/exam, and the incident will be submitted to the Department of Computer

Science for further review. Note that honor code violations can result in a permanent notation being placed on the student's transcript. Evidence of cheating may include a student being unable to satisfactorily answer questions asked by the instructor about a submitted solution. Cheating includes not only receiving unauthorized assistance, but also giving unauthorized assistance. For class files kept in Unix space, students are expected to use Unix file permission protections (chmod) to keep other students from accessing the files. Failure to adequately protect files may result in a student being held responsible for giving unauthorized assistance, even if not directly aware of it.

Students may still provide legitimate assistance to one another. Students should avoid discussions of solutions to ongoing assignments and should not, under any circumstances, show or share code solutions for an ongoing assignment. All students are responsible for knowing the rules. If you are unclear about whether a certain activity is allowed or not, please contact the instructor.

## 7 Class Schedule

**Day One:**

- The first part of the bootcamp provides the essential foundation of R needed to grasp the concept of R and increase your comfort level to explore some of the concepts and tasks further on your own.
- We will cover practical issues in preforming analysis which includes programming in R, reading data into R, accessing R packages, debugging, and organizing and commenting R code and R notebook.
- Beside covering basic R concepts and language fundamentals, we will introduce key concepts like tidy data and related "tidyverse" tools.
- Import data from csv and web
- Use dplyr package from "tidyverse" toolbox.

**Day Two:**

- During Day II, we take the gloves off and expand on concepts of tidy data (data manipulation), processing and manipulation of complex and large datasets, handling textual data, and basic data science tasks.
- Managing dataframes
- Cover the essential exploratory techniques for summarizing data, creating new data, and apply grammar of graphics concept
- Upon completing this part, you will have fluency at the R console and will be able to create tidy datasets from a wide range of possible data sources. (flat files, web)