

**STA5075Z,2023**

**Data Science Masters Statistical Computing**

**30 January 2023 to 10 February 2023**

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. In this section, we cover some of the basics of how R works and how to use it via RStudio. We will also look at RMarkdown as a powerful tool for integrating the data analysis workflow into presentations of different form.

### Concepts & Topics for this section

1. Getting set up and running with R on your computer
2. Basic R concepts and syntax
3. Basic data structures and calculations in R
4. Structure of vectors in R and calculations on them
5. R Markdown

**Learning Objectives:** By the end of this session you should:

- Have R and RStudio installed and running on your computer
- Be able to carry out basic calculations in R
- Be able to carry out calculations with vectors
- Know where to find help and additional information on aspects of R
- Understand the basic idea of R Markdown

### Activities:

1. **Read Chapters 1 and 2 of the notes**
2. **Run all the code** in the notes and figure out what it does
3. **Read up** on various aspects of R by following the links provided in the notes
4. **Solve Prac 1** and **submit your R code via the 'Prac 1' assignment tab by Tuesday, 31 January 2023 at 23:59 pm.**

In this section, we will look at some of the main data objects used in R.

### Concepts & Topics for this section

1. Working with matrices in R
2. Working with data frames in R

### Learning Objectives

By the end of this session you should:

- Be able to set up matrices in R
- Be able to do basic matrix algebra in R
- Understand the structure of data frames
- Know how to subset data frames and get summary statistics on their variables

### Activities:

1. **Read Chapter 3 of the notes.**
2. **Run all the code** in the notes and figure out what it does
3. **Solve Prac 2** and **submit your R code via the 'Prac 2' assignment by Wednesday, 01 February 2023 at 23:59 pm**
4. **Solve Prac 3** and **submit your R code via the 'Prac 3' assignment by Wednesday, 01 February 2023 at 23:59 pm.**

This section is about visualising data and producing simple graphics in R.

### Concepts & Topics for this section

1. Producing effective graphics in R
2. Using these graphics in documents

### Learning Objectives

By the end of this session you should:

- Be able to create simple graphics (including scatter plots, histograms, box plots and a few others) in R
- Be able to customise these graphics
- Be able to add to graphics using lower-level plotting functions
- Be able to save the graphics in various format using graphics device functions
- Be able to use these graphics in documents using R Markdown

### Activities:

1. **Read Chapter 4 of the notes**
2. Run all the code in the notes and figure out what it does
3. Complete **Prac 4** and [submit your R code via the Prac 4 assignment tab by Thursday, 02 February 2023 at 23:59 pm](#)
4. Complete **Prac 5** and [submit your R code via the Prac 5 assignment tab by Thursday, 02 February 2023 at 23:59 pm](#)
5. Complete **Prac 6** and [submit your R code via the Prac 6 assignment tab by Thursday, 02 February 2023 at 23:59 pm.](#)

In this section, we cover methods for reading data into R and for writing data from R to a file.

### Concepts & Topics for this section

1. Write data to file
2. Read data from some of the most common file types

### Learning Objectives

By the end of this session you should:

- Be able to read in data from text-based data files with arbitrary formatting
- Be able to read in data from Excel (.xls and .xlsx) files and be able to handle common problems with this file type
- Be able to read in data from simple data bases and data files stored on the internet
- Be able to set up a transparent and reproducible workflow when reading in data
- Have some tricks up your sleeve to deal with the common problems encountered when reading data into R

### Activities:

1. **Read Chapter 5 of the notes**
2. **Run all the code** in the notes and figure out what it does
3. **Complete** Prac 7 **and submit your R code via the Prac 7 assignment tab by Thursday, 02 February 2023 at 23:59 pm.**

In this section, we explore how programming works in R.

### Concepts & Topics for this section

1. Loops
2. Conditional expression
3. Conditional loops
4. Functions
5. Vectorization

### Learning Objectives

By the end of this session you should:

- Be able to use 'for' loops and conditional loops to program R to carry out operations repeatedly
- Be able to write your own functions in R
- Be able to use 'if - else' statements
- Understand the concept of vectorisation
- Be able to use the 'apply()' and related functions

### Activities:

1. **Read Chapter 6 of the notes**
2. **Run all the code** in the notes and figure out what it does
3. **Complete Prac 8 and** submit your R code via the 'Prac 8' assignment tab by Friday, 03 February 2023 at 23:59 pm.
4. **Complete Prac 9 and** submit your R code via the 'Prac 9' assignment tab by Friday, 03 February 2023 at 23:59 pm.
5. **Complete Prac 10 and** submit your R code via the 'Prac 10' assignment tab by Friday, 03 February 2023 at 23:59 pm.
6. **Complete Prac 11 and** submit your R code via the 'Prac 11' assignment tab by Friday, 03 February 2023 at 23:59 pm.

## **5. Wrap-up of Week 1 and Intro to Week 2**

**Week 1 – Day 5 (03-Feb-2023)**

- Additional (Optional) Exercises
- Additional Readings

In this section we begin to apply the knowledge and programming skills gained in Week 1 to tackle a variety of problems in R

### Concepts & Topics for this section

1. Slot machine example & problems
2. Snakes & Ladders problem

### Learning Objectives

- Understand how to approach solving problems via programming
- Understand the importance of formulating a problem before you attempt to solve it
- Practice formulating problems and discussing them with your team mates
- Understand that there is most often more than one way to solve a problem
- Be able to write your own functions in developing solutions to problems
- Understand the importance of commenting your code as a means of explaining your logic and process to someone else

### Activities

1. **Read Chapter 7 of the notes**
2. **Run all the code** in the notes and ensure you understand what it does
3. **Complete Prac 12** and submit your R code via the 'Prac 12' assignment tab by Tuesday, 07 February 2023 at 23:59 pm.
4. **Complete Prac 13** and submit your R code via the 'Prac 13' assignment tab by Tuesday, 07 February 2023 at 23:59 pm.



In this section we will introduce Monte Carlo Simulation and use it in tackling a variety of problems

### Concepts & Topics for this section

1. Monte Carlo Simulation
2. Statistical distributions (Uniform, Normal, Exponential)
3. Random samples
4. Integration
5. Cumulative distribution function (CDF)
6. Distribution of the sampling mean
7. Random sums

### Learning Objectives

- Develop a good understanding of Monte Carlo simulation
- Be able to apply Monte Carlo methods to solve a variety of problems
- Firm up your understanding of important statistical concepts (distributions, random samples, simulation, distribution of the sampling mean, etc.)

### Activities

1. **Read Chapter 8 of the notes**
2. **Run all the code** in the notes and ensure you understand what it does
3. Complete **Prac 14** and **submit your .Rmd and HTML files via the Assignments tab by Wednesday, 08 February 2023 at 23:59 pm**

In this section we will introduce a few optimisation algorithms and investigate applying them in R

### Concepts & Topics for this section

1. Optimisation
2. Local vs. global optima
3. Gradient descent
4. Newton's method
5. Convergence criteria
6. Optimisation in R using `optim()`
7. Maximum likelihood estimation

### Learning Objectives

- Develop a good understanding of optimisation and some optimisation methods
- Be able to conduct optimisation in R to solve a variety of problems
- Further grow your understanding of the optimisation methods covered in this tutorial by conducting your own further research

### Activities

1. **Read Chapter 9 of the notes**
2. **Run all the code** in the notes and ensure you understand what it does
3. **Complete Prac 15** and [submit your .Rmd and HTML files via the Assignments tab by Thursday, 09 February 2023 at 23:55 pm](#)

This section will not cover any more theory. You are required to Complete Prac 16 and submit your .Rmd and HTML files under the Assignments tab.

### Concepts & Topics for this section

1. Prac 16

### Learning Objectives

- Be able to conduct optimisation in R to solve a variety of problems
- Further grow your understanding of the optimisation methods covered in this practical by conducting your own further research
- Collaborate with your team mates

### Activities

1. **Read Chapter 10 of the notes**
2. **Complete Prac 16** and submit your .Rmd and HTML files via the Assignments tab by Friday, 10 February 2023 at 23:59 pm.

In this section we will introduce you to parallel and high performance computing.

### Concepts & Topics for this section

1. Parallel computing
2. High performance computing
3. UCT HPC
4. Cloud Computing
5. AWS EC2

### Learning Objectives

- Be able to conduct parallel computing in R
- Further grow your understanding of the parallel computing methods covered in this practical by conducting your own further research
- Familiarise yourself with the various R packages that enable parallel computing
- Grow your understanding of logistic regression and cross-validation
- Collaborate with your team mates!

### Activities

1. **Read Chapter 11 of the notes**
2. **Conduct your own research** on logistic regression for classification and cross-validation
3. **Complete Prac 17** and **submit your .Rmd and HTML files via the Assignments tab by Monday, 13 February 2023 at 23:59 pm**