**Introduction to R**



Anish Mahapatra

Table of Contents

[1. Introduction 3](#_Toc120382011)

[2. What is Data? 4](#_Toc120382012)

[3. What is R? 4](#_Toc120382013)

[4. The applications of R - (Talk about RStudio) 5](#_Toc120382014)

[5. What is an R Script? 7](#_Toc120382015)

[6. What are packages in R 8](#_Toc120382016)

[7. Commonly used packages in R 8](#_Toc120382017)

[8. R for statistical analysis 8](#_Toc120382018)

[9. Excel versus R versus Python 8](#_Toc120382019)

[10. Data Types in R 8](#_Toc120382020)

[11. Coding with R – Installation of R / R Studio 8](#_Toc120382021)

[12. How to read a dataset in R 8](#_Toc120382022)

[13. Data Operations in R 9](#_Toc120382023)

[14. Data Wrangling in R 10](#_Toc120382024)

[15. Data Wrangling in R 10](#_Toc120382025)

[16. Visualization with R 10](#_Toc120382026)

[17. Applications of Machine Learning in R 10](#_Toc120382027)

[18. AI in R 10](#_Toc120382028)

[19. Thank you! 10](#_Toc120382029)

[20. Author 12](#_Toc120382030)

[21. Archive 12](#_Toc120382031)

# Introduction

Hello everyone!

Welcome to the course on the Introduction to Data Science with R. The very fact that you registered for this course proves that you’re headed a step in the right direction to start your journey and to enhance your position in the Data science domain.

But before we begin let me tell you what you can expect from this course.

We will primarily be speaking about data science, some of the applications, not all of them since it’s a field with a vast scope across different domains. I will also be introducing you to the statistical coding language **R** which acts as a great tool to actively compute complex statistical analysis in a few simple steps.

The objective that I would like to achieve is to make sure you leave this course a little better versed in R and it’s uses without delving too deep into the technical side of things.

We will be dealing with questions such as What is data? What is R? What are packages? And a whole lot more that will give you a basic understanding of the beautiful field and how to go about it in the right way.

My name is Anish Mahapatra. I’m a Senior Data Scientist with a MSc degree in Data science with over than 5 years of work experience. I currently work at a Fortune 200 Retail Giant and I also am a consultant that’s worked with multiple Fortune 500 clients in the data science domain. A great way to know more about me would be to simple Google my name – Anish Mahapatra.

Without furthor ado, let us now dive into the fundamentals of Data Science and R!

# What is Data?

It’s a simple question, right? What is Data? Data is a collection of discrete values that convey information. This can be in the form of quantity, quality, fact, statistics, and other basic units of meaning.

The world is driven by data – all the way from your mobile phones, e-commerce, banking and even your cars!

When looking at it from an uber perspective, it is important to understand what are companies spending their money on. The reason Data Science has such high holding for the next couple of decades is that companies are in the process of leveraging their data.

Data can be seen as rows and columns, where we are able to get contextualized information in a concise manner. It can be somethingas simple as your shopping list.

Now, how does this work in the digital world? We can have data in the form of csv files, excel files, a database and on the cloud platform. All of this data flows from different sources and there is great value in understanding what all of it means.

It might be easy to understand data that comprises of 10-20 rows. But, what if we increase that number to a 100 million rows? It becomes difficult to “understand” the data then, because it does not even open in Excel.

This is where a language like R can help. It can help you understand, infer and visualize large volumes of data.

Let’s have a look at What is R in our next session. This is probably a great place to tell you to not fear code. You have gotten this far and it is an achievement, and we will make this as easy to understand as possible. Stick with me.

# What is R?

R is a great language for everyone to learn as it’s sole purpose is to collect, analyse, interpret and present data. In fact, this is the very definition of statistical analysis.

R is a programming language and environment for statistical computing and graphics to analyze and visualize data.

Every language has associated rules, called syntax that govern it. There are three items that are a part of the syntax of R are:

1. **Variables**: Objects that can store data
2. **Comments**: To make the code easier to understand and more readable
3. **Keywords**: Words that are reserved for the compiler of the language. What is a compiler? It converts instructions into machine-level language.

So, why do people actually use R?

As one of the top five programming languages currently. R is used by large companies in disciplines such as Fintech, Retail, Social Media, Healthcare etc. to do the following:

* Data Analytics
* Statstical Inference
* Machine Learning

There is a good reason why a language becomes one of the most widely used in the world. Some of the reasons for R programming to be one of the op languages are:

* **Open-source free language**: R and it’s suite of tools is completely free to download and use
* **Platform Independent**: Whether you have a mac, windows, linux or even a cloud server, R can be run on all platforms with ease
* **Leading tool for machine learning**: R programming is one of the leading tools for machine learning, statstics and data analysis. It is one of the most requested language in the Data Science job market.
* **Flexibility**: R can be integrated with many other languages, and this makes it even more useful, irrespective of the language that is used
* **Community Support**: Any questions that you have with R can easily be answered with a Google search. This is because there are a lot of people that support the community.

Alright Anish, you seem to have sold me on the potential of R, not just in the job market, but also in the search for great Data Science jobs. So, how do we actually use it?

I’m glad you asked. In the next session, we will talk more about some of the applications of R, before we deep-dive and look at how you can get started in under ten minutes!

# The applications of R - (Talk about RStudio)

Hi everyone, welcome back to this session, where we will discuss the applications of R. At any point if the complexity of data increases, we need ways that we can interpret, interface and analyze the data. There are use-cases in multiple domains such as:

* **Research & Academics**: Since R Programming is used primarily for statistical computing, it is a great resource for students & teachers to leverage
* **Information Technology (IT) Sector**: The IT sector uses R to enable businesses to build statistical, computational tools, data handling and business intelligence.   
    
  There are use-cases such as the analysis of web activity, recommendation systems, marketing campaign analysis and even matchmaking systems
* **Banking & Finance**: Risk management of assets, providing loans to potential customers, upselling financial & banking products are some use-cases where R can be tremendously helpful.   
     
  It can also be used for fraud detection, loan stress modelling, loan stress test simulation and a lot of other forms of risk analytics.
* **E-Commerce**: E-commerce uses the potential of R to work on use-cases such as marketing mix modelling, cross-selling, targeted advertising, sales and financial data modelling.
* **Healthcare**: R is greatly used in healthcare for a plethora of operations. It can be used to do analysis in the fiels of Genertics, bioinformatics, epidemiology, and, drug discovery. It can also be used to understand the statstical understanding of the success of clinical trials.

Alright, great, it’s good to know that it is used in a lot of verticals across multiple industries. Now, the question is are we getting closer to understanding how can I get started with R? Yes, we are getting close, but we are not just there yet.

An IDE or integrated development environment is a software to build code, which combines developer tools into a single interface. Luckily for us, the IDE for R, called RStudio is a all-in-one solution to quickly get started with R!

In the next section, we will learn more about the core concept of how we will set out to build fantastic things with R – using R scripts!

# What is an R Script?

Welcome to this session, where we will understand more about R Scripts! Put Simply, R Scripts are just text files that contain commands to run in R. They form the playground with which we can let R work it’s magic.

Let’s discuss the importance of scripts in R, here are some of the reasons why scripts are important in R:

* **Helps reproduce your work**: R scripts enable you to keep a systematic record of your analysis, so that you can reproduce your work later or collaborate with other users
* **Do not need to remember what you did**: Keeping a record of your R code on R Scripts means that you do not need to rely on your memory to figure out what you did
* **Allows you to add comments**: R Scripts enables you to add comments to your code, sothat you not only remember what you did, but also **why** you did it
* **Can be run in batch mode**: R scripts can be run in batches, which means a series of scripts can be run parallelly or in a series
* **Can add inputs in R scripts**: R programming allows you to give inputs to your program, so that it will be easier to generate the outputs from the inputs so that it is easier to reconstruct your code while showcasing it
* **Can be combined with markdown**: One of the biggest advantages of R scripts is that it lets you combine R with a language called ‘markdown’, which let’s you generate fully reproducible documents that have both the inputs and the outputs.   
    
  This is done using the knitr and rmarkdown packages!

“Packages” New Terminology Alert! We will learn more about packages in our upcoming session on “What are packages in R?”, and we will also get to see examples of packages and their super powers.

Meanwhile, I’d like to quickly touch upon RStudio. RStudio has it’s own text editor, which you can use to write R Scripts. Later in this course, we will teach you how to download RStudio and write your first R Script with it! See you in the next session!

# What are packages in R

Welcome to this session, where we we will learn more about packages in R. Let’s say we would like to calculate the sum of a set of numbers in R. We can either write the R code from scratch or we can use a function.

For a lot of simple or complex tasks, the community has already written the code. Wouldn’t it be great if we could all leverage code that has already been written?

Great news! This already exists. Packages in R are a collection of R functions, data and compiled data in a well-defined format, that is created to add specific functionality.

One of the reasons that R is one of the top 5 programming langauges in the world is because it has over 10,000+ user contributed packages and this number is growing constantly! This is a great test of how supportive the open-source community is.

This enables anyone working in R to simply install a package from the package repository and use it per their use-case. CRAN, or Comprehensive R Archive Network contains an archive of the latest and previous version of R, the documentation and all the versions of the R packages.

Now that we have a clear understanding on packages in R, in the next session we will understand more about the commonly used packages in R.

# Commonly used packages in R

Welcome back to this session, where we will learn about some of the popular packages in R. Some of the more popular packages that are used in R are:

* **dplyr**: The dplyr package is used to perform data analysis and data transformations. Where there are several functions to work with dataframes
* **ggplot2**: ggplot is the mostpopular package to create graphics. It enables the user to create elegant plots and graphs from the data
* **tidyr**: tidyr is a package that is used to tidy the data. In R, a data is considered tidy when each row represents an observation, and each variable represents a column
* **shiny**: R goes beyond simple visualizations. R Shiny is an interactive web application that lets the user to embed visualization in a web platform and also has in-built support for animation, graphs, plots and charts
* **caret**: Caret stands for Classification and Regression training. The Caret package helps model complex regression and classification problems. CaretEnsemble can also be used to combine (ensemble) different models

Some of the more popular packages you can look up are tidyverse, plotly, knitr and mlr3. Now that we know what packages in R are let’s move on to on of R’s primary purpose – Statistical Analysis. Let’s learn more in the next session.

# R for statistical analysis

There is imeense scope in R for statistical analysis. In this session, we will briefly cover some ways that we can use R for statistical analysis. Of course, to understand the topic in depth, we would need more time.

We will touch upon the main pointers through which we can use R for statistical analysis:

* **Descriptive analysis using R**
  + Descriptive analysis is the process of using current and historical data to identify trends and relationships. It is the simplest form of analysis.
  + We can understand how to measure central tendency of the data through Mean, Median and Mode with R
  + The variation in the data can also be understood through Range, IQR, which stands for Inter-Quartile Range and Deviation.   
      
    It’s fine if you are not completely aware of the terminology. This is 10th Grade mathematics and a simple revision through the course will help you understand the possibilities of R can help understand the underlying trends in the data.
* **Data visualization with R** 
  + R enables visualization by giving the user the ability to plot the data in terms of bar charts, scatter plots, pie charts, histograms and along with many other visuals
  + We can aslo take subsets of the data, and modify detaisl in the plot such as the axis name, the colors, the markings and denotions, text size and color and even plot multiple charts in a single space
  + R also enables us to plot time series plots, where we are able to see trends over a period of time. We will have a chance to go through some of the visualizations in R in a later section.
* **Probability with R**: We can deep-dive in probability with the help of R Programming as well, where concepts such as union and interesection, factorial, and permutation and combination are covered.
* **Probability Distributions** 
  + This is a large segment to cover. Essentially, R is one of the top programming langauges used to cover the probability distributions that can be derived from the data.
  + It covers concepts such as Central Limit Theorem. Normal Probability Distribution, Binomial Probability Distribution, Poisson Distribution and many more.
* **Inferential Statistics**: This is a bit of an advanced topic, but a fundamental understanding of Inferential statistics can help us understand the potential power.   
    
  Inferential Statistics is a field of statistics that uses analytical tools to draw conclusions about a poplation by analysing random samples.It can cover the following concepts:
  + **Hypothesis testing**: This covers concepts such as Z-test, F-test, T-test, ANOVA test etc.
  + **Regression Analysis**: Linear Regression, Nominal Regression, Logistic Regression, Ordinal Regression etc.

We have now successfully gone over some of the key statistical abilities of R.

# Excel versus R versus Python

Welcome to this session, where we will get to learn more about R versus Python. We will also touch upon Excel and discuss it as all you may have used Excel at some point.

Excel

# Data Types in R

# Coding with R – Installation of R / R Studio

# How to read a dataset in R

This can also be done manually in Rstudio by choosing the import dataset feature found in the file menu.

# Data Operations in R



# Data Wrangling in R

# Data Wrangling in R

# Visualization with R

# Applications of Machine Learning in R

# AI in R

# Thank you!



# Author

**Anish Mahapatra**   
*Senior Data Scientist*   
LinkedIn: [https://www.linkedin.com/in/anishmahapatra](about:blank) ([LinkedIn](about:blank))   
Email: anishmahapatra01@gmail.com

# Archive

1. Introduction
2. What is Data?
3. What is R?
4. The applications of R - (Talk about RStudio)
5. What is an R Script?
6. What are packages in R
7. Commonly used packages in R
8. R for statistical analysis
9. Excel versus R versus Python
10. Data Types in R
11. Coding with R – Installation of R / R Studio
12. How to read a dataset in R
13. Data Operations in R
14. Data Wrangling in R
15. Visualization with R
16. Applications of Machine Learning in R
17. AI in R
18. Thank you!