Data Science

Introduction to Data Science

Agenda

01 Need of Data Science

02 Life Cycle of Data Science

03 Applications of Data Science

04 Introduction to R

05 Introduction to R-Studio

06 Data Types & Operators in R

Data

Data

Well, it's just a collection of facts!

-0.879

348

"My name is Sam"

$A = \pi r^2$

0 1

1 0

Data Back Then

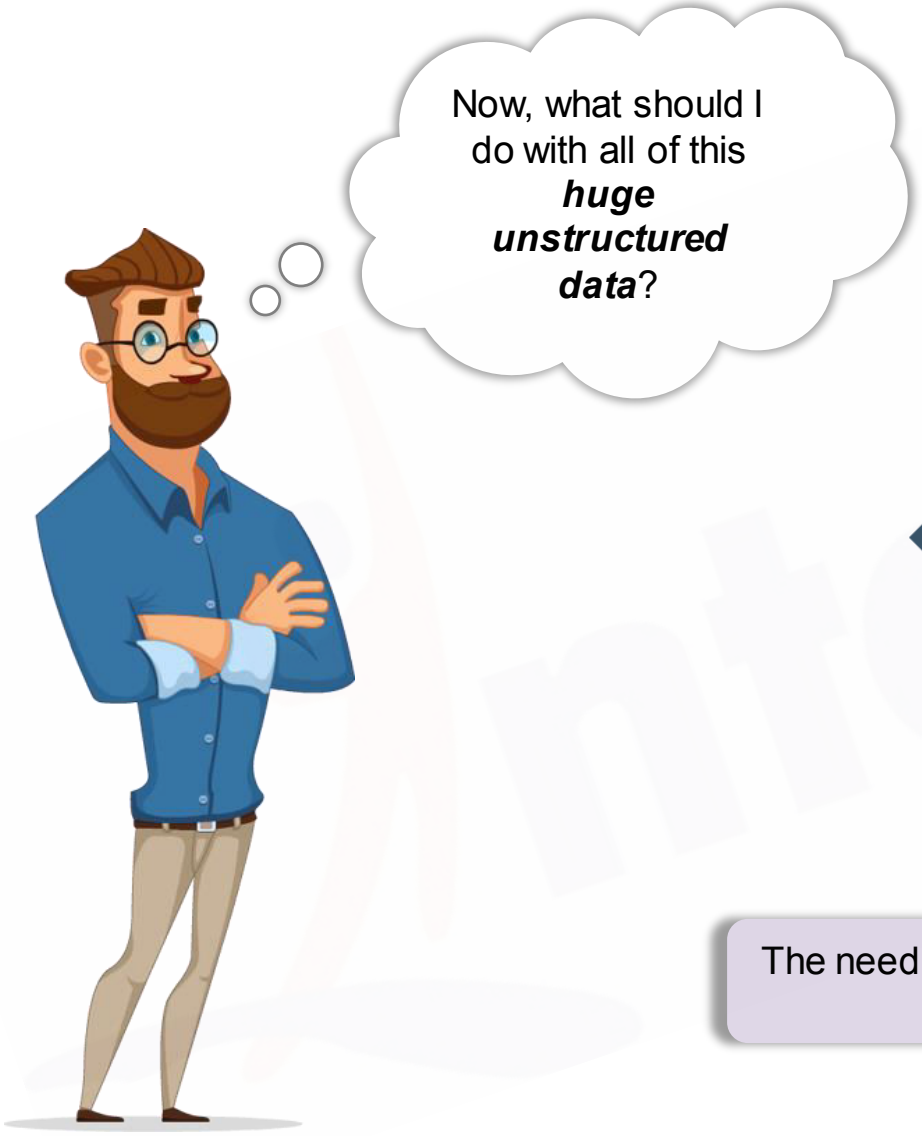
Data Today

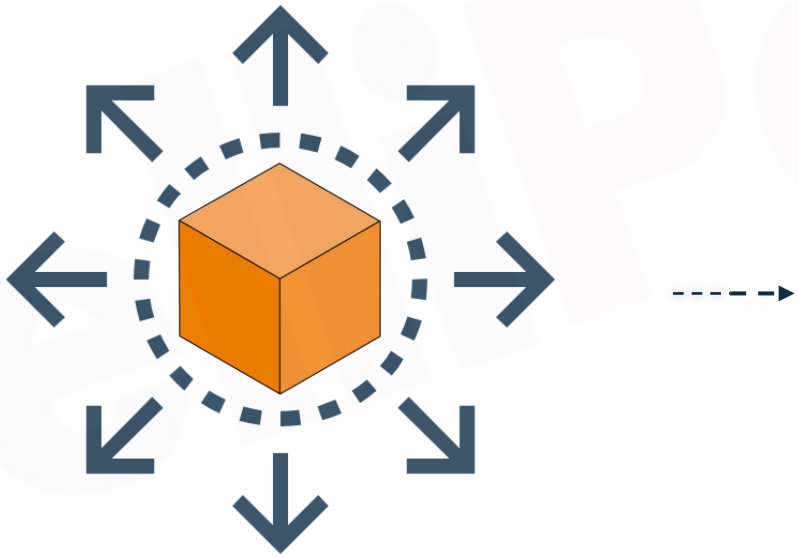
Unstructured

Multiple Formats

Humongous Sized

Need of Data Science





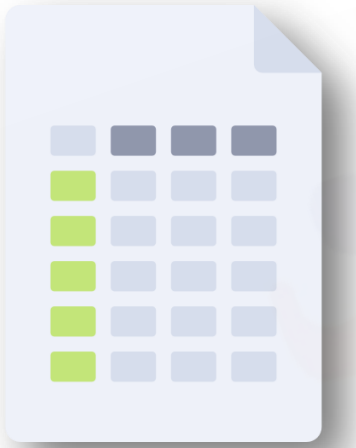
The need to understand and analyze data to make better decisions is what gave birth to Data Science.

Need of Data Science

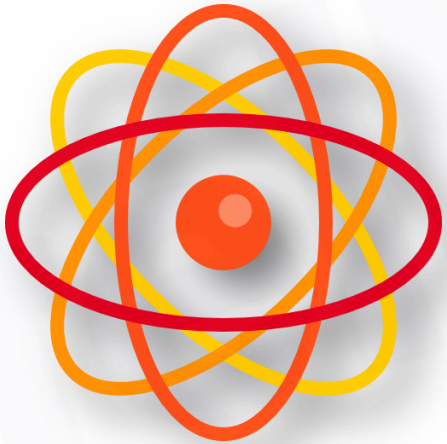
- Understand the data
- Find interesting insights
- Make informed decisions

What is Data Science?

Applying science on data to make the data talk to us



Data



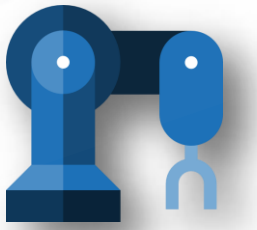
Science

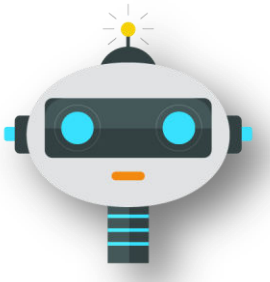
Data Science

What is Data Science?

Data Science is an umbrella term which encompasses multiple domains.

- Data Visualization
- Data Manipulation
- Statistical Analysis
- Machine Learning





Types of Data Analytics

Prescriptive

Comprehensive, accurate and effective visualization

Predictive

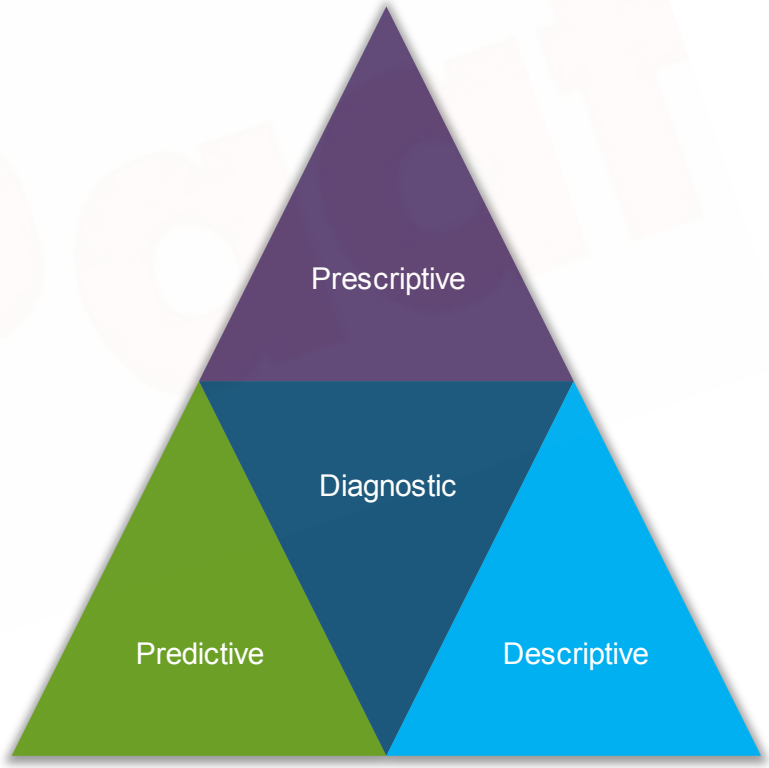
Ability to drill down to the root cause

Diagnostic

Historical patterns being used to predict specific outcomes using algorithms

Descriptive

Applying advanced analytical algorithms to make specific recommendations and strategies

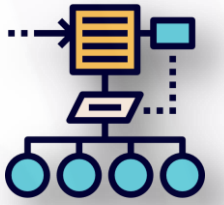


Life Cycle of Data Science

Life Cycle of Data Science

Data Acquisition

Data Preprocessing



Model Building

Pattern Evaluation

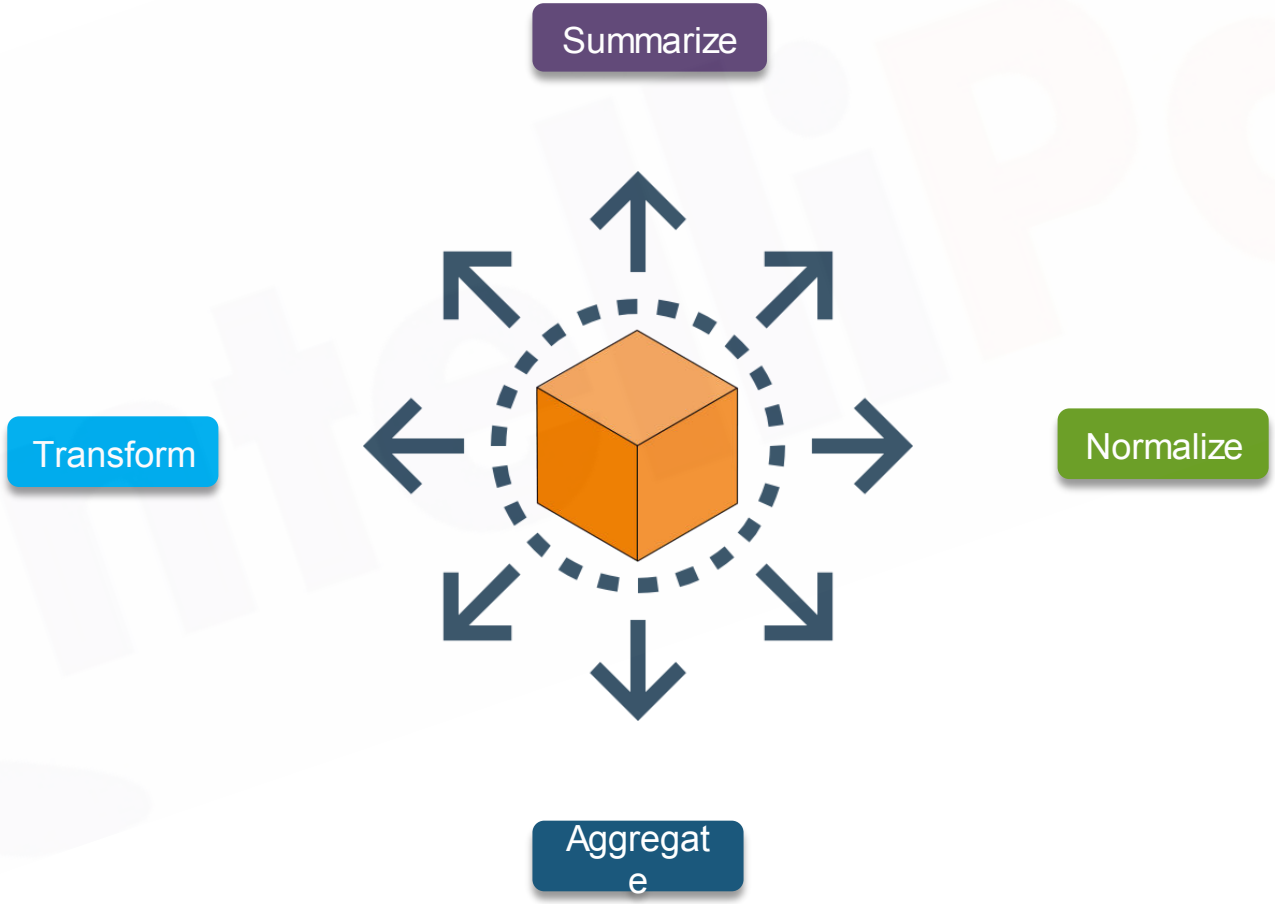
Knowledge Representation

Data Acquisition

Data comes from multiple sources and is present in multiple formats. This data has to be integrated and stored in one single location.

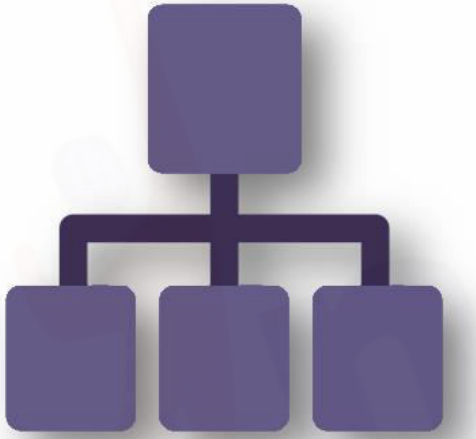
Data Preprocessing

Once data acquisition is done, the raw data has to be processed to bring it to the right format.



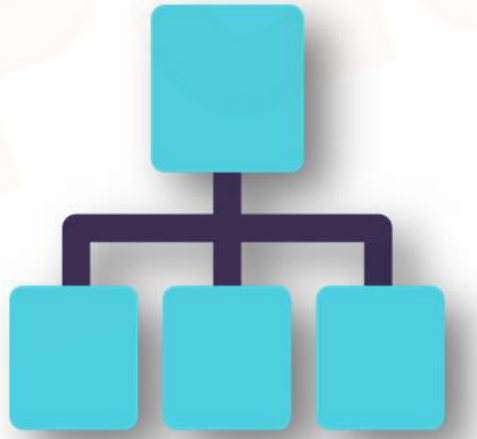
Model Building

Model building is the process where we apply different scientific algorithms to find interesting insights from the data.



Linear Regression

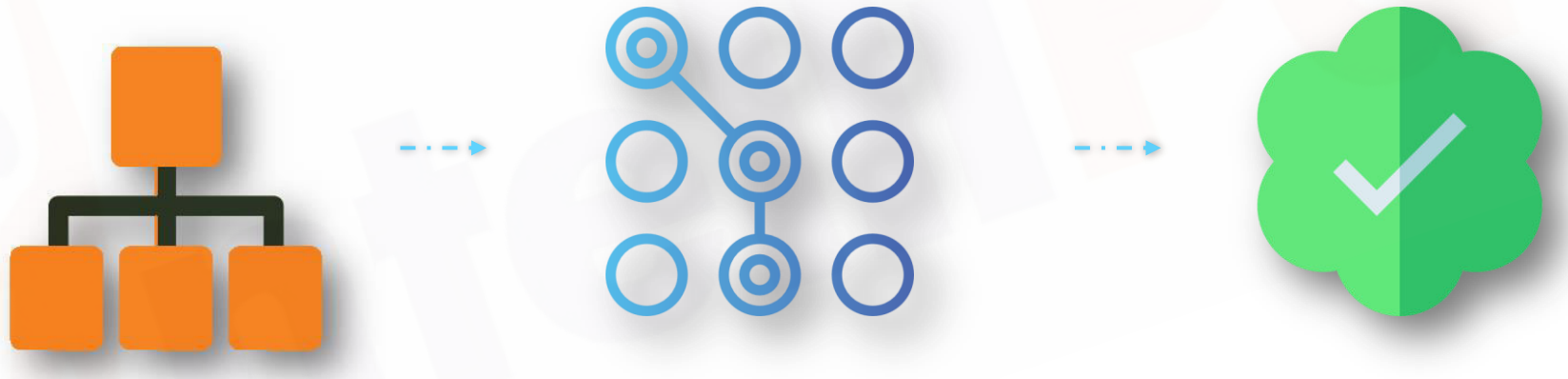
K-Means



Random Forest

Pattern Evaluation

The model gives us some patterns/information. These patterns have to be evaluated, i.e., here we have to evaluate whether the obtained information is new, correct and useful.



Model

Pattern

Evaluation

Knowledge Representation

Once the information is validated, it can be represented with simple aesthetic graphs.

Application of Data Science in Different Industries

Application of Data Science in Telecom

- Analytical Customer Relationship Management (ACRM)
- Fraud Reduction
- Bad Debt Reduction
- Price Optimization

Application of Data Science in Banking

- Acquire and Retain Customers
- Detect Fraud
- Improve Risk Control
- Optimize Product and Portfolio Model

Application of Data Science in E-commerce

- Enhance Customer Engagement
- Customize Offers and Promotions
- Maintain Effective Supply Chain Management
- Improve User Experience

Introduction to R

Introduction to R

R is a language for data analysis and statistical analysis.

Introduction to R

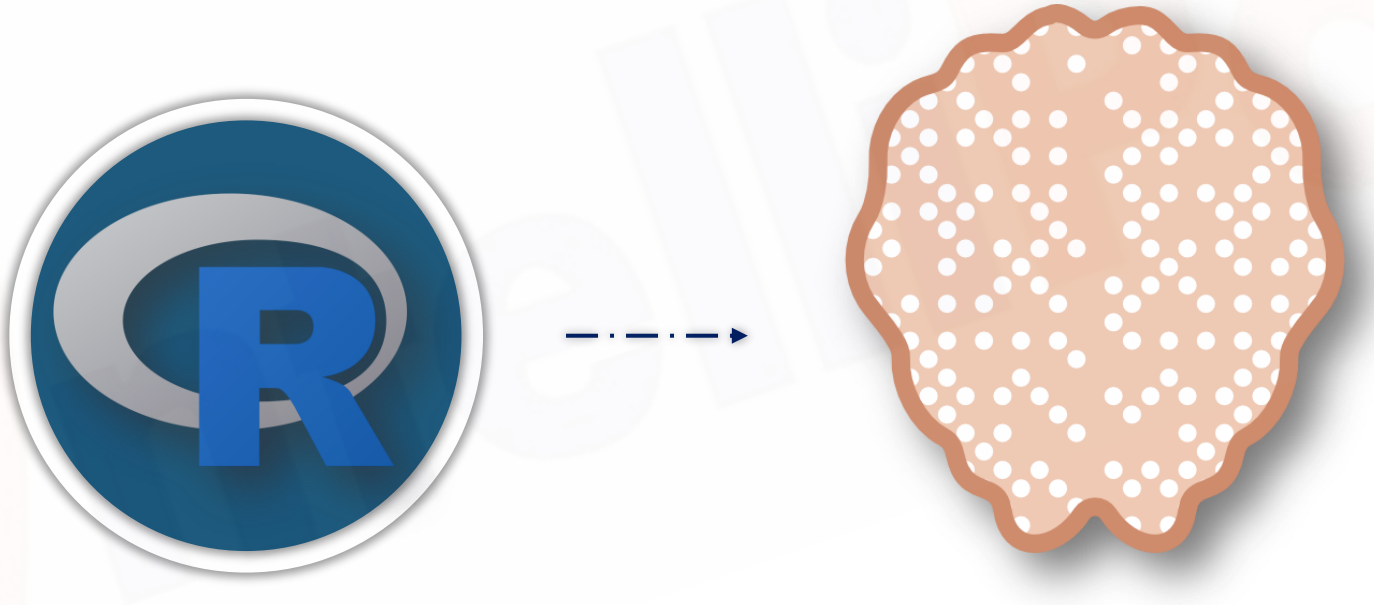
R is a visualization tool.

Introduction to R

R is an open-source, cross-platform compatible software.

Introduction to R

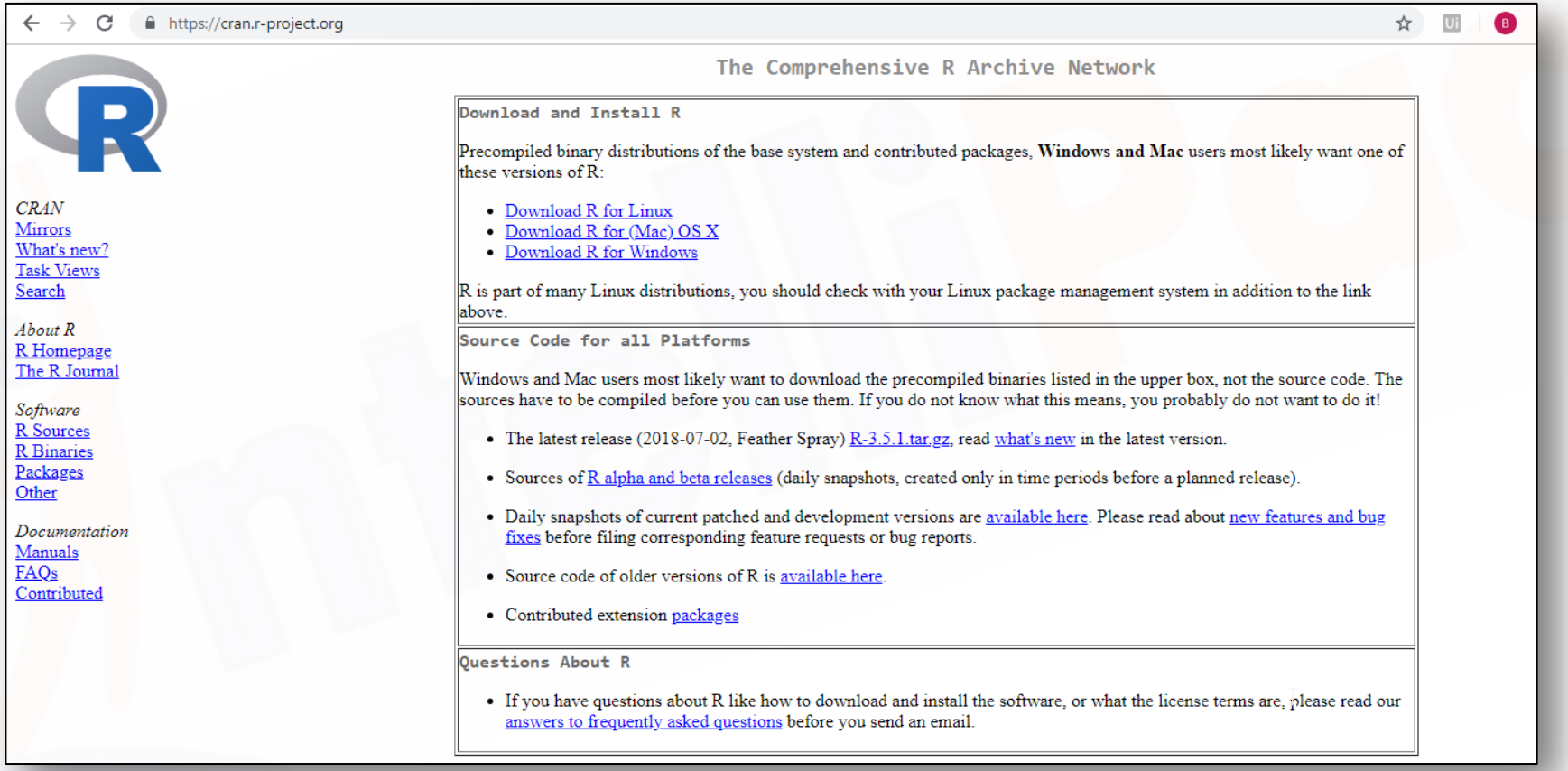
R is a Turing complete language.



Installing R

Installing R

You can install R from https://cran.r-project.org/



R-Studio

R-Studio

R-Studio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.

Setting Working Directory

Change your working directory with the setwd() function, such as:

```
setwd("~/mydirectory")
```

Note that slashes always have to be *forward* slashes, even if you're on a Windows system.

For Windows, the command might look something like: → `setwd("C:/Sham/Documents/RProjects")`

Customizing R-Studio

R-Studio options are accessible from the Options dialog, **Tools > Options** menu (**R-Studio > Preferences** on a Mac) and include the following categories:

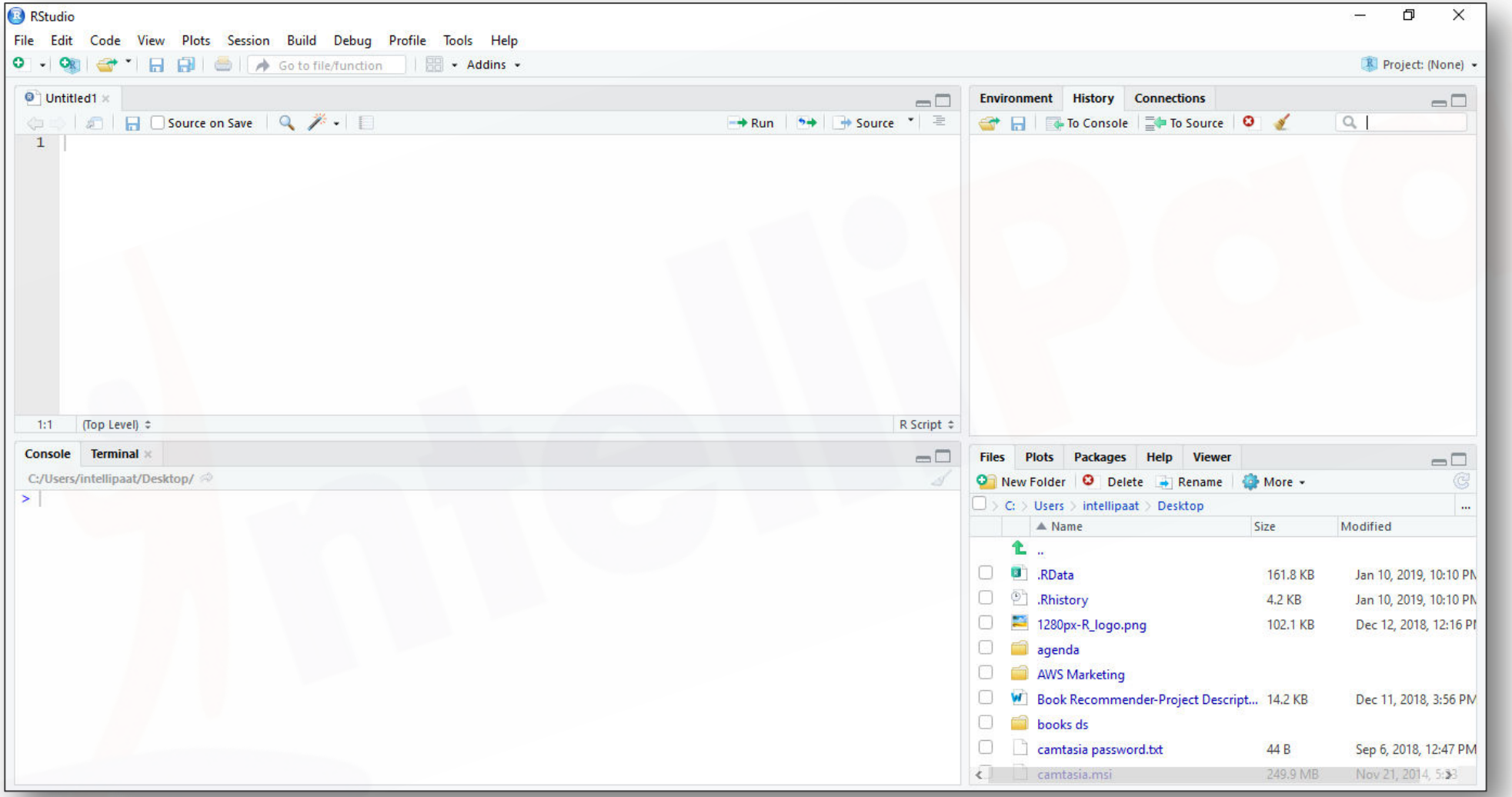
General R Options	Default CRAN mirror, initial working directory, workspace and history behavior
Source Code Editing	Enable/disable line numbers, selected word and line highlighting, soft-wrapping for R files, parent matching, right margin display, console syntax highlighting, configure tab spacing and set default text encoding
Appearance & Themes	Specify the font size and visual theme for the console and source editor
Pane Layout	Locations of console, source editor and tab panes; set which tabs are included in each pane

Customizing R-Studio

Packages	Set default CRAN repository and specify package development options
Sweave	Configure Sweave compiling options and PDF previewing
Spelling	Choose main dictionary language and specify spell checking options
Git/SVN	Configure locations of Git and SVN binaries and create and/or view SSH RSA keys
Publishing	Enable publishing apps and documents from IDE and set account

R-Studio GUI

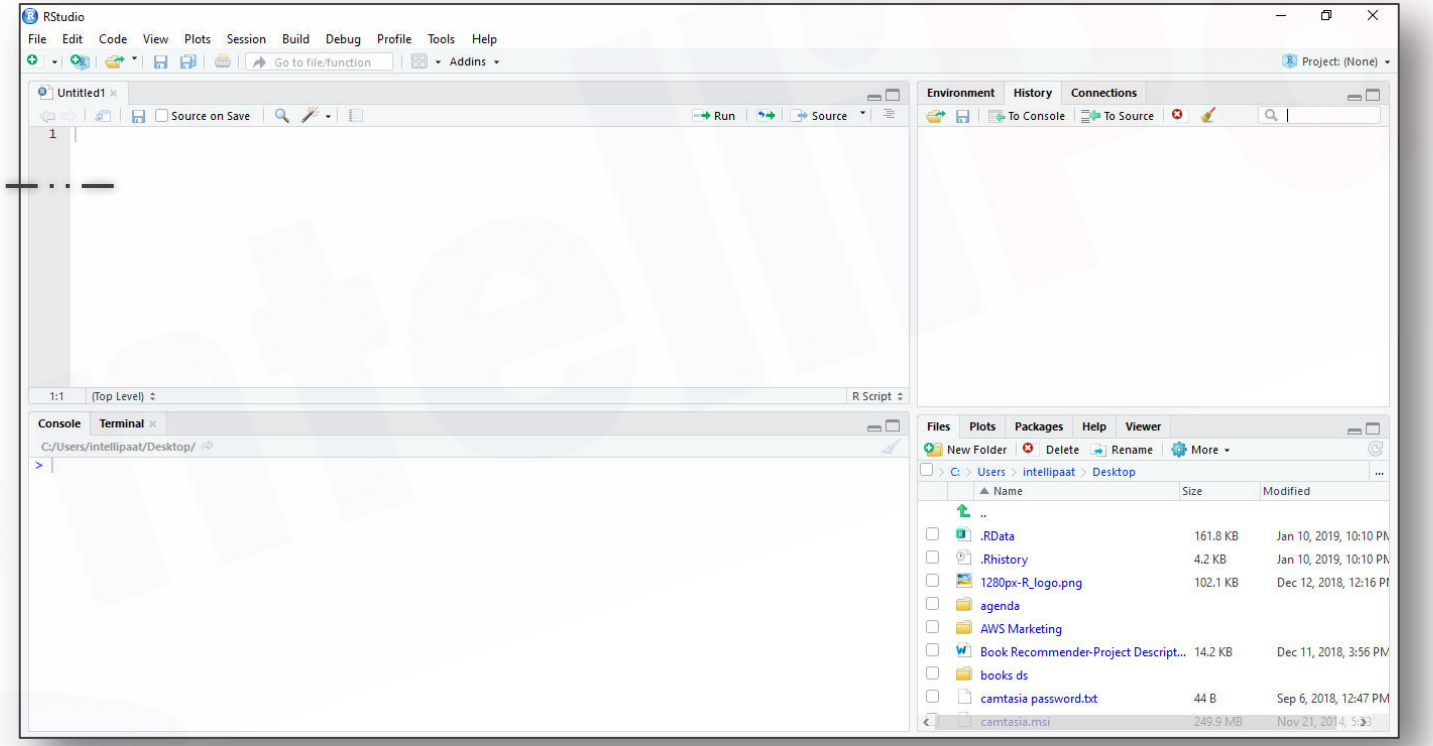
R-Studio GUI



R-Studio GUI

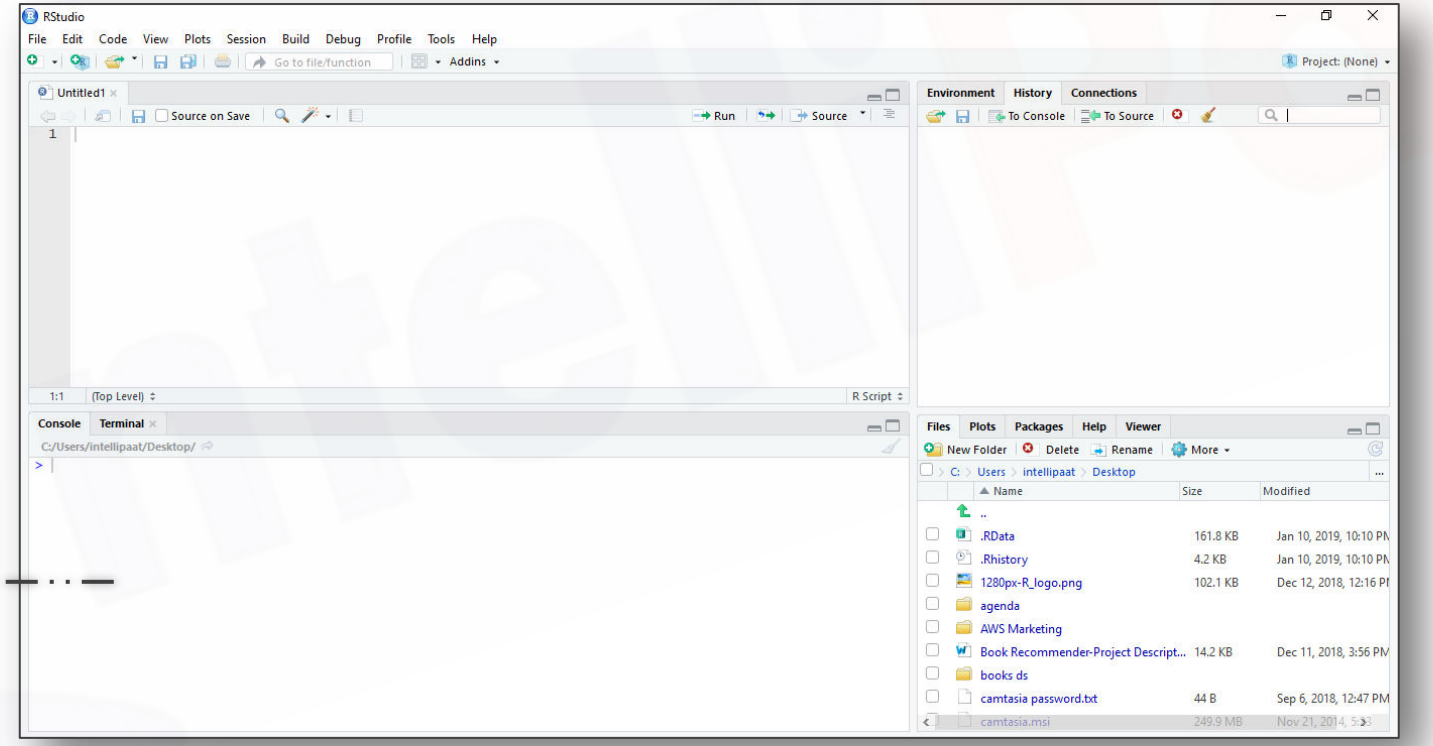
In the top left corner of the screen, one can see a script editor window. Within this pane, one can edit his or her R script.

Script Window



R-Studio GUI

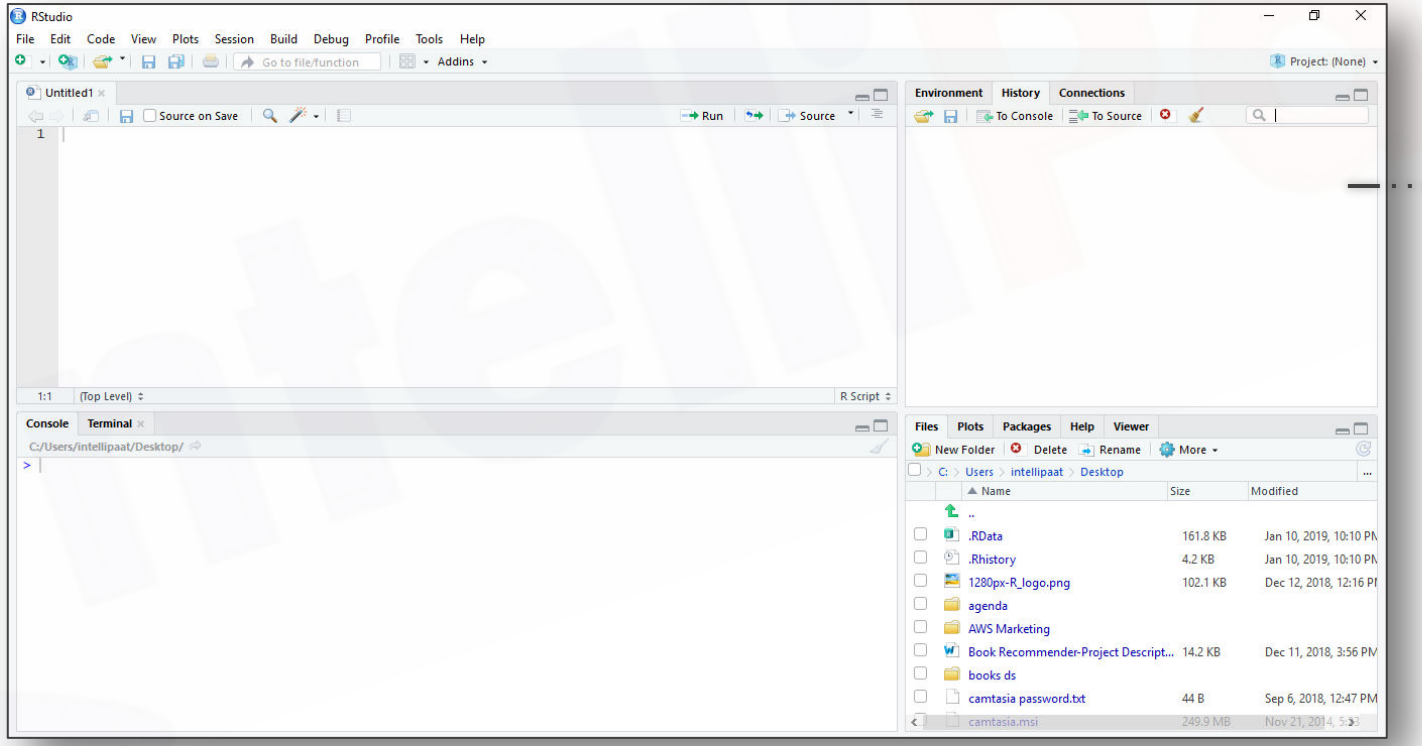
Results of the script execution, together with the script lines that generated these results, will be displayed in the Console window located in the bottom left corner of the screen.



Console Window

R-Studio GUI

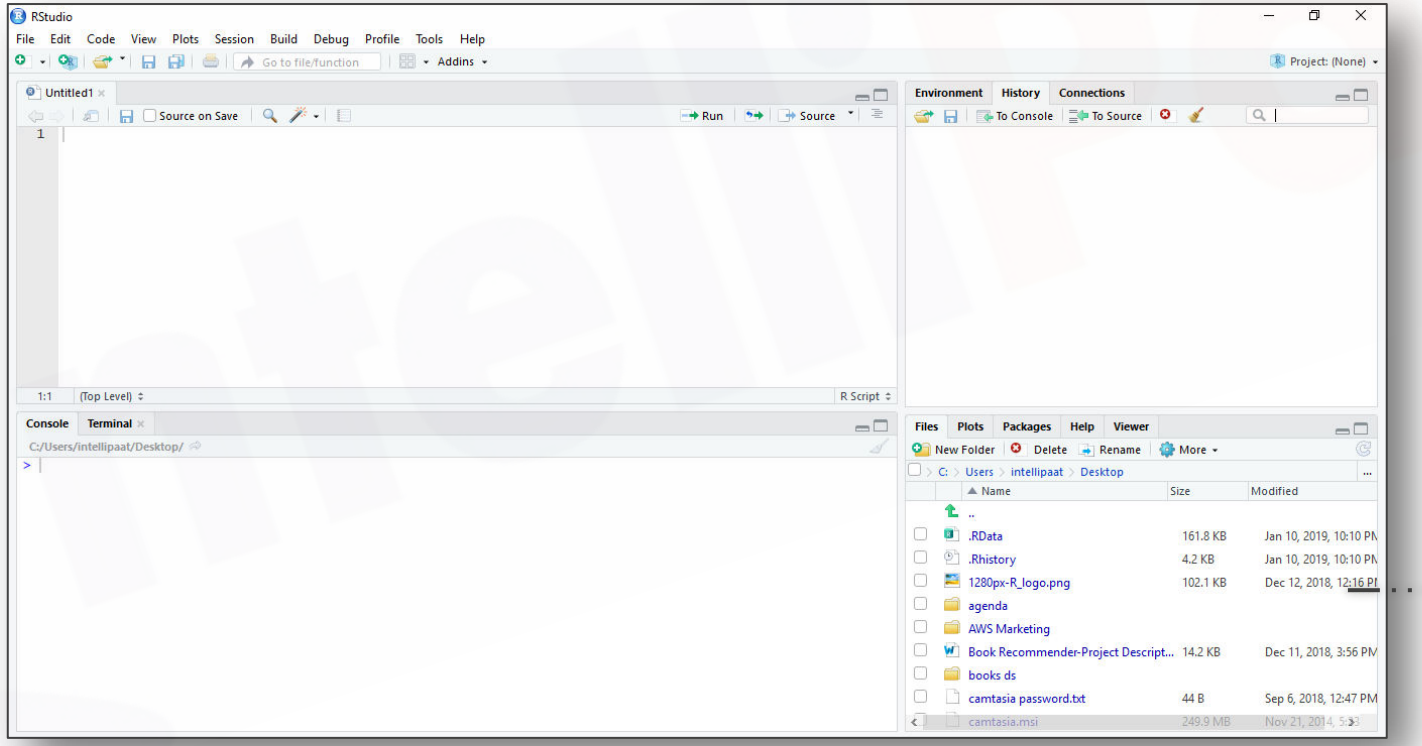
The top right pane of the screen provides information about the variables and data structures used or generated by the script. This is the so-called “Environment” window.



Environment Window

R-Studio GUI

The window on the bottom right corner of the screen shows information about the files and packages used by the project and allows one to view plots (or visualizations) generated by R and also access help for various elements of R syntax.



Plots/Help

R Packages

R Packages

Packages are collections of R functions, data and compiled code in a well-defined format.

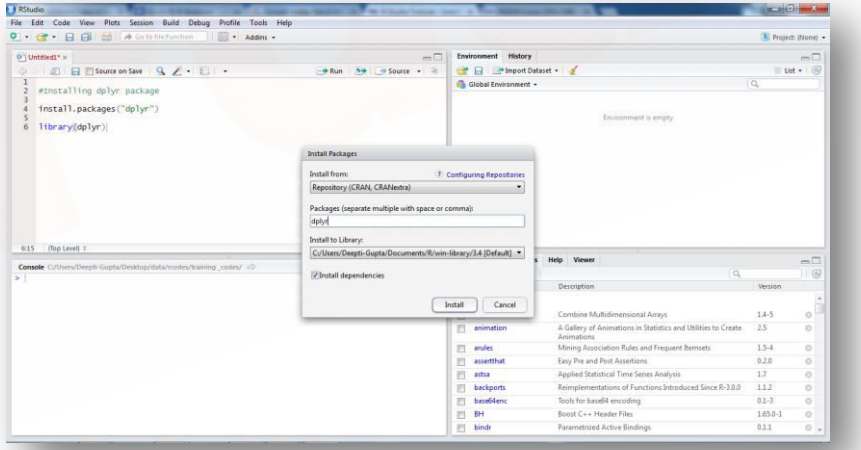
The directory where packages are stored is called the library.

R comes with a standard set of packages. Others are available for download and installation.

R Package	Function
.libPaths()	# Get library location
library()	# See all packages installed
search()	# See packages currently loaded
detach("package:pkg")	# Unload the loaded package
Install.packages("package")	# Install the package
library("package")	# Load the package
library(help= "package")	# List package contents

Steps to Install R Packages

1. Run R-Studio
2. Click on the Packages tab in the bottom-right section and then click on Install. The following dialog box will appear.
3. In the Install Packages dialog, write the package name you want to install under the Packages field and then click Install. This will install the package you searched for or give you a list of matching packages based on your package text.



Getting Help with R

Getting Help with R

The help() function and ? help operator in R provide access to the documentation pages for R functions, data sets and other objects, both for packages in the standard R distribution and for contributed packages.

The help() function can be used to access information about a package in your library—for example, help(package="MASS")—which displays an index of available help pages for the package, along with some other information.

Help Command	Function
help.start ()	# General help
help(lm)	# Help about function lm
example(lm)	# Show an example of function lm
help(package)	# List help page for “package”
?package	# short form for “help(package)”

Variables in R

Variables in R

A variable is a temporary storage space where you can keep changing values.

Variable

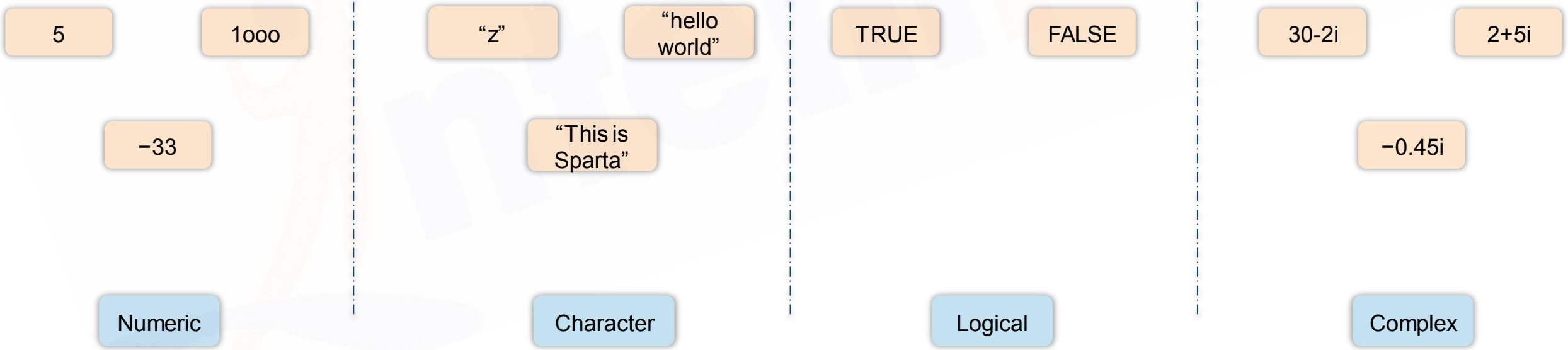
Variable

Variable

Data Types in R

Data Types in R

Every variable is associated with a data type.



Operators in R

Operators in R

Operators help in performing certain manipulations on top of the data and variables.

- Assignment Operators
- Arithmetic Operators
- Relational Operators
- Logical Operators

Assignment Operators

Assignment operators are used to assign a value to an object.

Operators

=

<-

->

Example

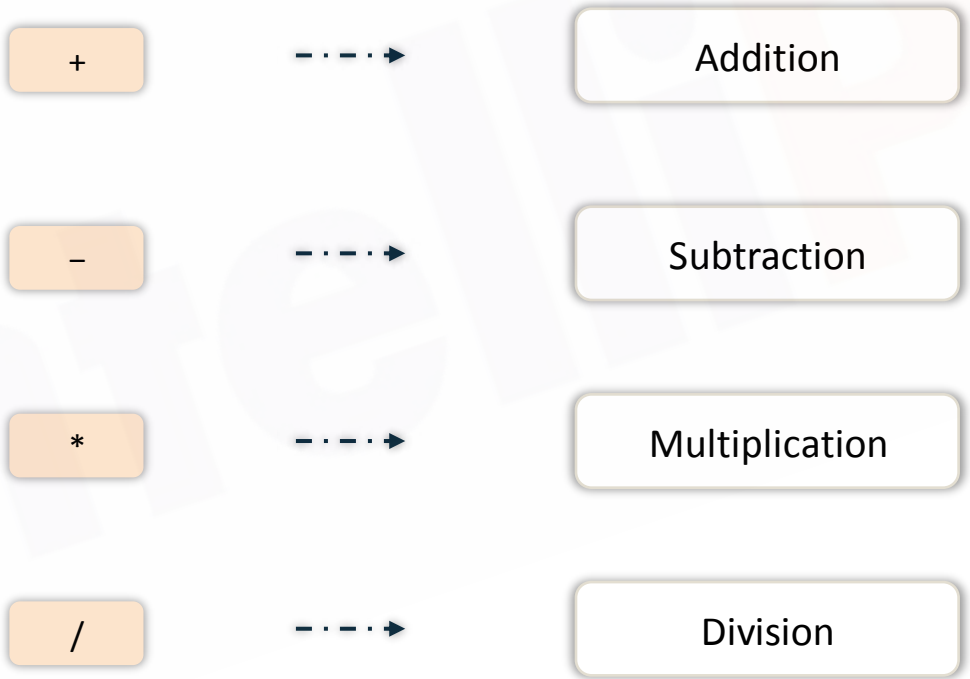
x = 10

y <- 20

30 -> z

Arithmetic Operators

Arithmetic operators are used to perform basic mathematical operations.



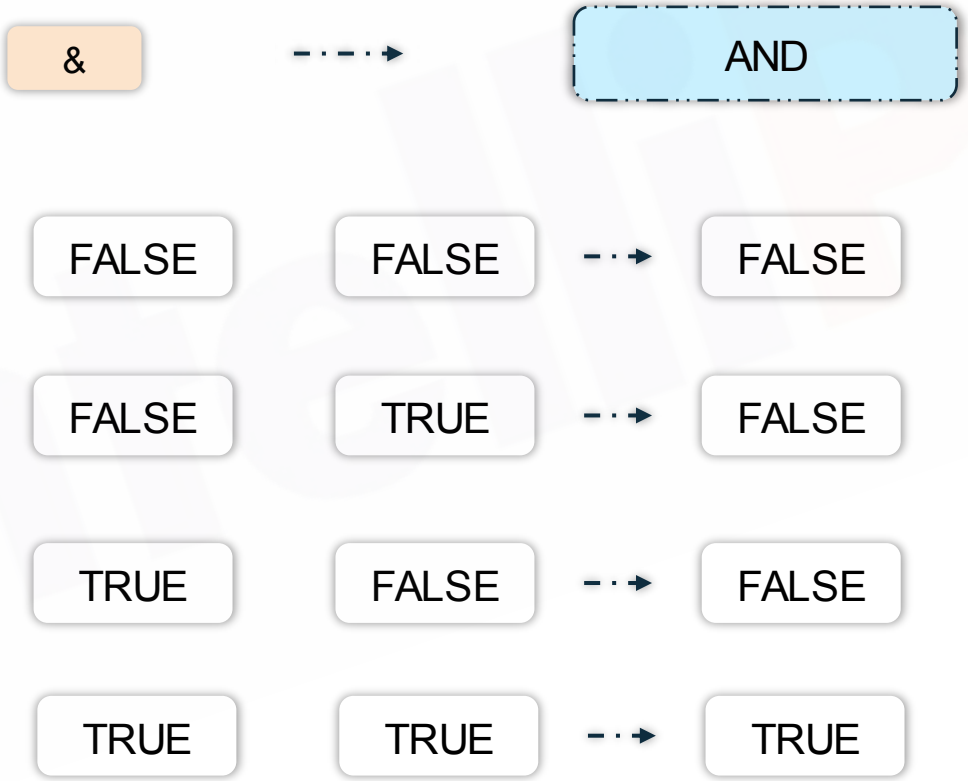
Relational Operators

Relational operators are used to test/define a relationship between two operands.

<	⇢	Less than
<=	⇢	Less than or equal to
>	⇢	Greater than
>=	⇢	Greater than or equal to
==	⇢	Is equal to
!=	⇢	Not equal to

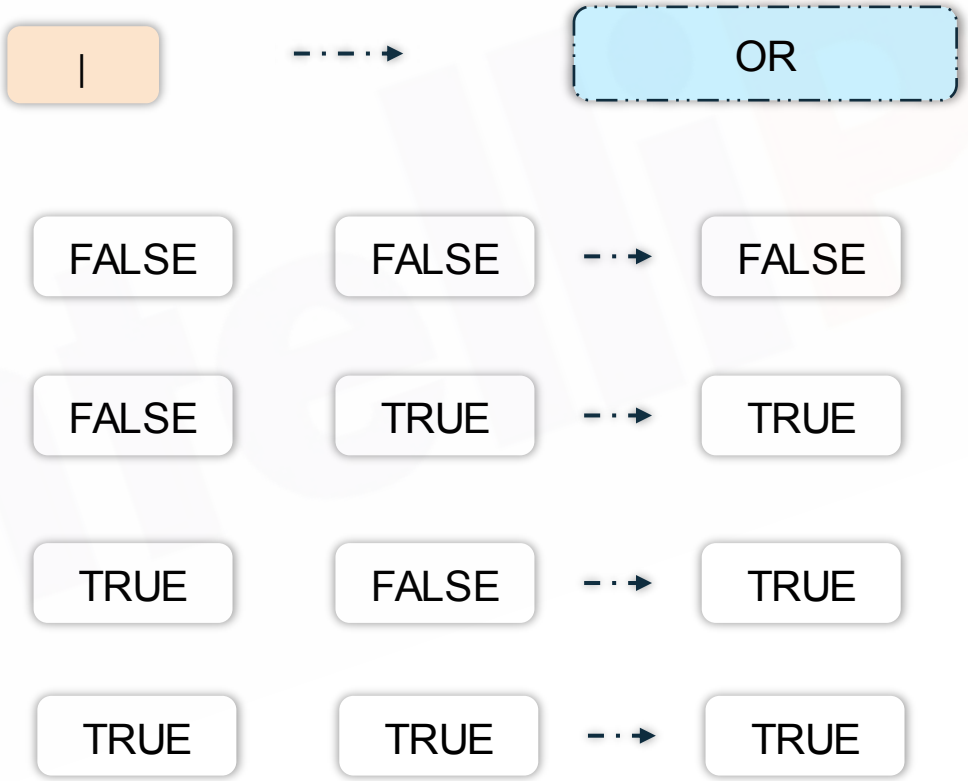
Logical Operators - AND

Logical operators are used to make a decision on the basis of a condition.



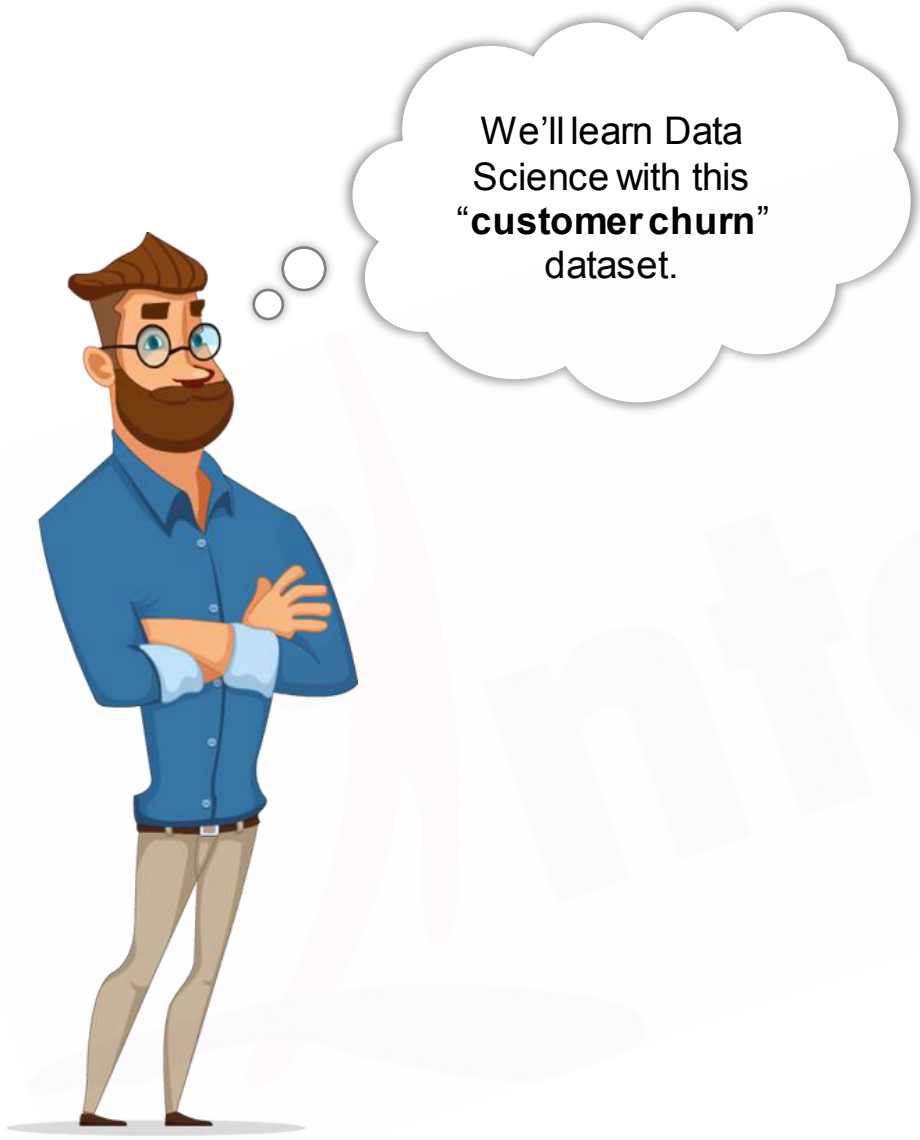
Logical Operators - OR

Logical operators are used to make a decision on the basis of a condition



Project-based Data Science Course

Data Science Project



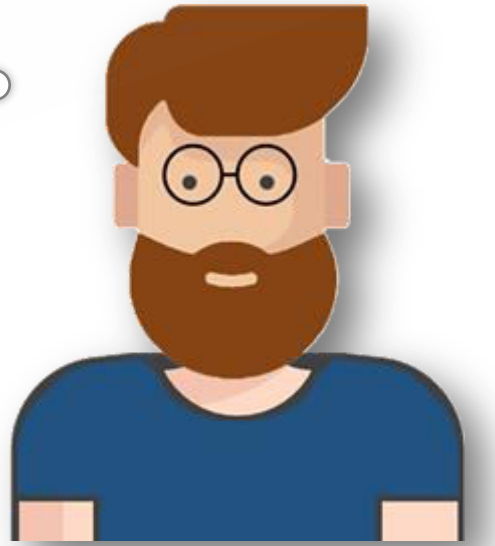
customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

Problem Statement

You are the Data Scientist at a telecom company "Neo" whose customers are churning out to its competitors. You have to analyse the data of your company and find insights.

Neo

I'll analyse my company's data completely to find why customers are churning out.



Tasks to be Performed

1

Data Manipulation

Find out hidden patterns in the “customer_churn” dataset by using apply family of functions and dplyr package

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

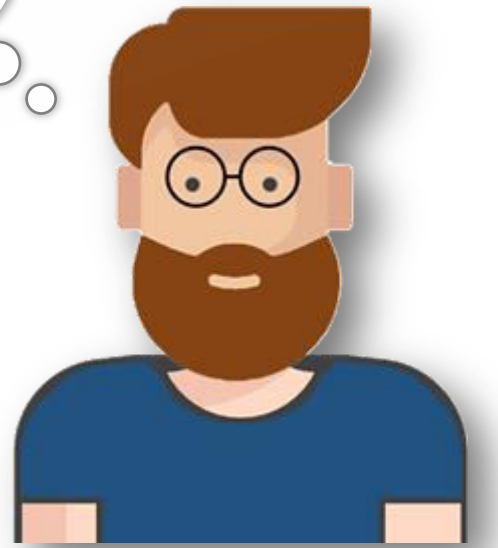
Tasks to be Performed

2 Data Visualization

Represent the data with graphs by using ggplot2 package

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

I'll depict the data pictorially to get a better understanding.



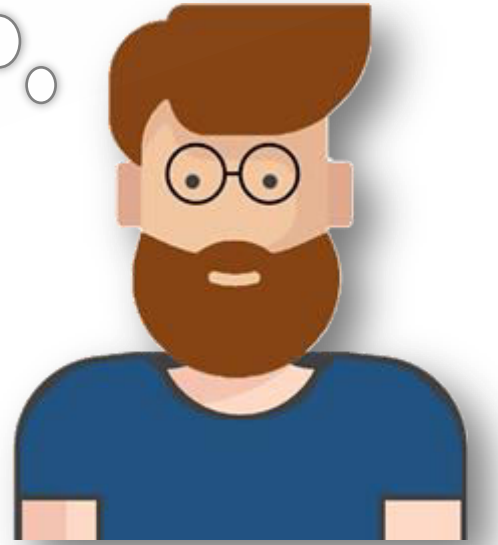
Tasks to be Performed

3 Linear Regression

Understand how the Monthly Charges of the customers vary with respect to other factors

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

I'll build a linear regression algorithm on top of the "customer_churn" data.



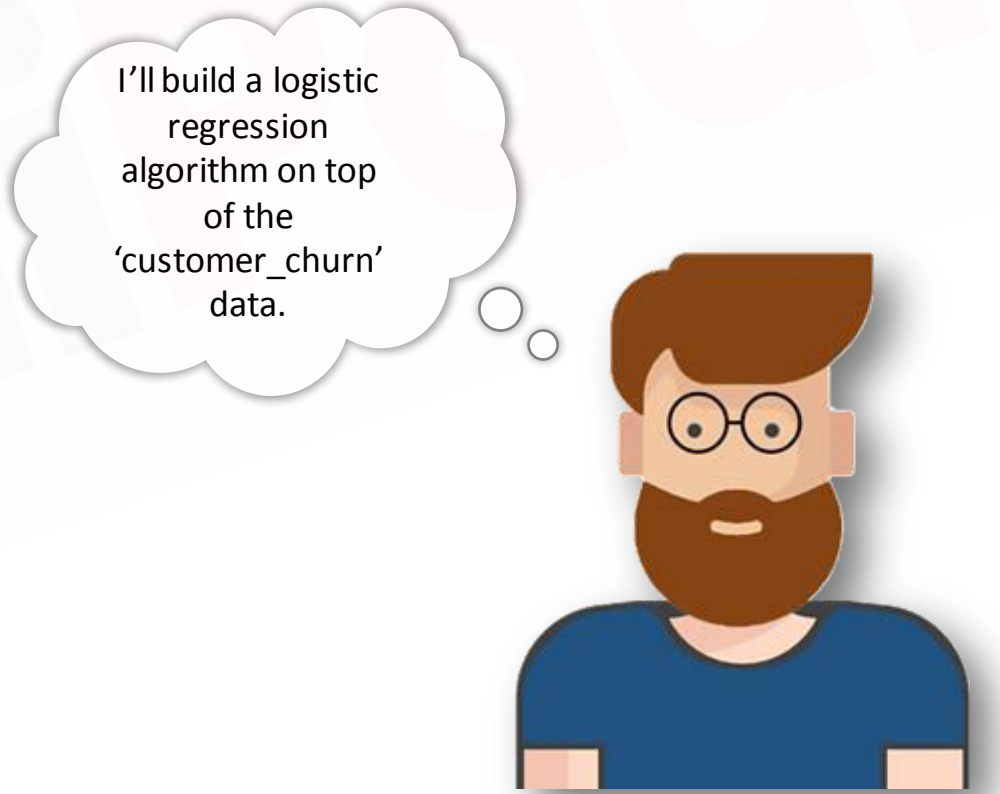
Tasks to be Performed

4

Logistic Regression

Get the probability of customers churning out with respect to other factors

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes



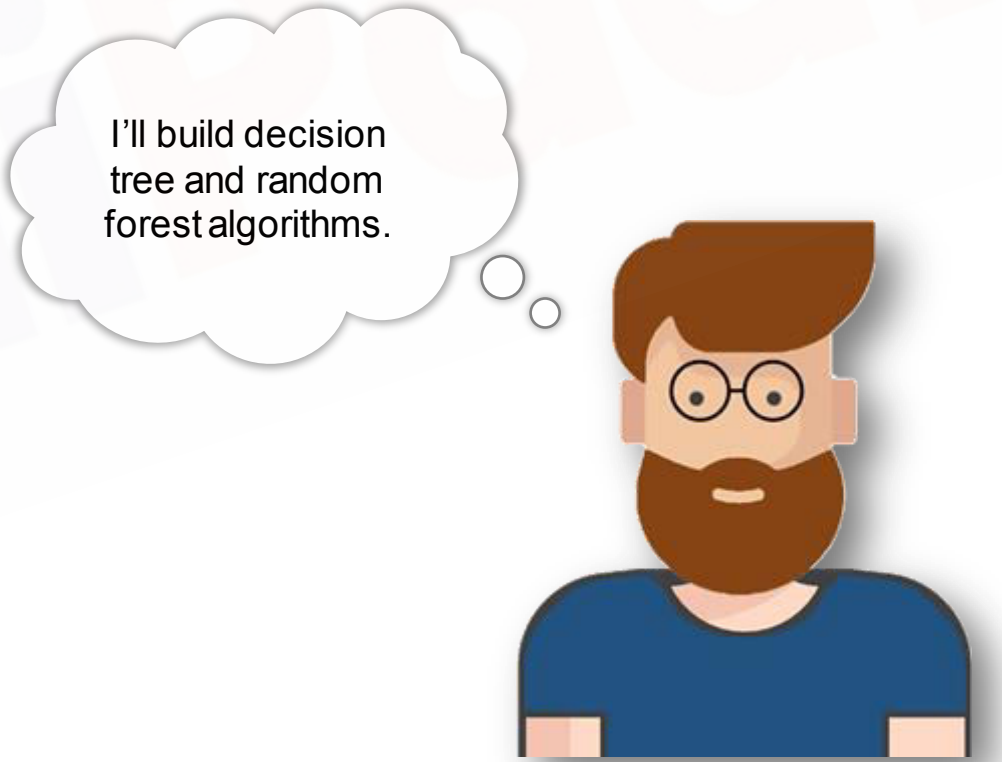
Tasks to be Performed

5

Decision Tree & Random Forest

Classify whether the customer will churn or not on the basis of other factors

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes



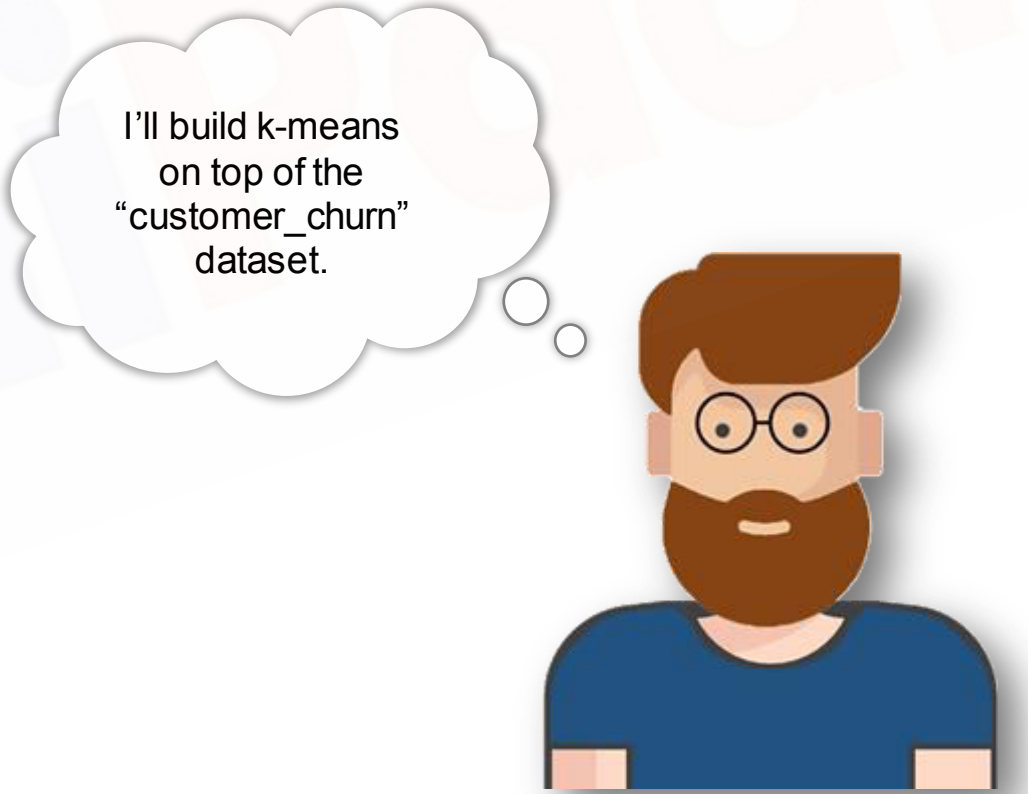
Tasks to be Performed

6

Clustering

Divide the customers into different clusters with k-means clustering

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes



Individual Modules

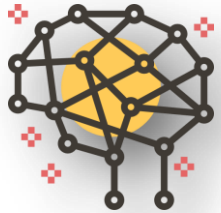
Individual Modules

Individual Modules which are not based on ‘customer_churn’ dataset

Market Basket Analysis

Recommendation Engine

Time Series



Deep Learning

IntelliPaat

Quiz

Quiz

Which of the following is not a type of analytics?

a. Descriptive

b. Business Intelligence

c. Predictive

d. Prescriptive

e. None of the above

Quiz

Which of the following is not a type of analytics?

Solution:

b. Business Intelligence

Quiz

Which of the following are the 4 Vs or dimensions of Big Data ?

a. Volume, Velocity, Variable & Vacuum

b. Volume, Velocity, Variety & Veracity

c. Volume, Vaccine, Variety & Variable

d. All of the above

Quiz

Which of the following are the 4 Vs or dimensions of Big Data ?

Solution:

b. Volume, Velocity, Variety & Veracity

Thank You

India : +91-7847955955

US : 1-800-216-8930 (TOLL FREE)

sales@intellipaat.com

24/7 Chat with Our Course Advisor