Introduction to Data Science

Year/Sem: III/I

UNIT - I Introduction: Definition of Data Science- Big Data and Data Science hype – and getting past the hype - Datafication - Current landscape of perspectives - Statistical Inference - Populations and samples - Statistical modeling, probability distributions, fitting a model – Over fitting. Basics of R: Introduction, R Environment Setup, Programming with R, Basic Data Types.

Definition of Data Science

- ✓ **Data Science** is an interdisciplinary field that focuses on extracting knowledge from data sets which are typically huge in amount. The field encompasses analysis, preparing data for analysis, and presenting findings to inform high-level decisions in an organization. As such, it incorporates skills from computer science, mathematics, statics, information visualization, graphic, and business.
- ✓ Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.
- ✓ It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.
- ✓ Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

In short, we can say that data science is all about:

- Asking the correct questions and analyzing the raw data.
- Modeling the data using various complex and efficient algorithms.
- Visualizing the data to get a better perspective.

- Understanding the data to make better decisions and finding the final result.

Example:

Let suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Note: Few important steps to help you work more successfully with data science projects:

- **Setting the research goal:** Understanding the business or activity that our data science project is part of is key to ensuring its success and the first phase of any sound data analytics project. Defining the what, the why, and the how of our project in a project charter is the foremost task. Now sit down to define a timeline and concrete key performance indicators and this is the essential first step to kick-start our data initiative!

- **Retrieving data:** Finding and getting access to the data needed in our project is the next step. Mixing and merging data from as many data sources as possible is what makes a data project great, so look as far as possible. This data is either

found within the company or retrieved from a third party. So, here are a few ways to get ourselves some usable data: connecting to a database, using API's or looking for open data.

- **Data preparation:** The next data science step is the dreaded data preparation process that typically takes up to 80% of the time dedicated to our data project. Checking and remediating data errors, enriching the data with data from other data sources, and transforming it into a suitable format for your models.

- **Data exploration:** Now that we have clean our data, it's time to manipulate it to get the most value out of it. Diving deeper into our data using descriptive statistics and visual techniques is how we explore our data. One example of that is to enrich our data by creating time-based features, such as: Extracting date components (month, hour, day of the week, week of the year, etc.), Calculating differences between date columns or Flagging national holidays. Another way of enriching data is by joining datasets — essentially, retrieving columns from one data-set or tab into a reference data-set.

- **Presentation and automation:** Presenting our results to the stakeholders and industrializing our analysis process for repetitive reuse and integration with other tools. When we are dealing with large volumes of data, visualization is the best way to explore and communicate our findings and is the next phase of our data analytics project.

- **Data modeling:** Using machine learning and statistical techniques is the step to further achieve our project goal and predict future trends. By working with clustering algorithms, we can build models to uncover trends in the data that were not distinguishable in graphs and stats. These create groups of similar events (or clusters) and more or less explicitly express what feature is decisive in these results.

Big Data and Data Science hype

- **Big Data:** This is a term related to extracting meaningful data by analyzing the huge amount of complex, variously formatted data generated at high speed, that cannot be handled, or processed by the traditional system.
- Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer.
- Big data is a buzzword used to describe immense volumes of data, both unstructured and structured, that can inundate a business on a day-to-day basis.
- Big data is used to analyze insights, which can lead to better decisions and strategic business moves

Source of Big Data:

- **Social Media:** Today's world a good percent of the total world population is engaged with social media like Facebook, WhatsApp, Twitter, YouTube, Instagram, etc. Each activity on such media like uploading a photo, or video, sending a message, making comment, putting like, etc create data.
- **A sensor placed in various places:** Sensor placed in various places of the city that gathers data on temperature, humidity, etc. A camera placed beside the road gather information about traffic condition and creates data. Security cameras placed in sensitive areas like airports, railway stations, and shopping malls create a lot of data.
- **Customer Satisfaction Feedback:** Customer feedback on the product or service of the various company on their website creates data. For Example, retail commercial sites like Amazon, Walmart, Flipkart, and Myntra gather customer feedback on the quality of their product and delivery time. Telecom companies, and other service provider organizations seek customer experience with their service. These create a lot of data.

- **IoT Appliance:** Electronic devices that are connected to the internet create data for their smart functionality, examples are a smart TV, smart washing machine, smart coffee machine, smart AC, etc. It is machine-generated data that are created by sensors kept in various devices. For Example, a Smart printing machine – is connected to the internet. A number of such printing machines connected to a network can transfer data within each other. So, if anyone loads a file copy in one printing machine, the system stores that file content, and another printing machine kept in another building or another floor can print out that file hard copy. Such data transfer between various printing machines generates data.
- **E-commerce:** In e-commerce transactions, business transactions, banking, and the stock market, lots of records stored are considered one of the sources of big data. Payments through credit cards, debit cards, or other electronic ways, all are kept recorded as data.
- **Global Positioning System (GPS):** GPS in the vehicle helps in monitoring the movement of the vehicle to shorten the path to a destination to cut fuel, and time consumption. This system creates huge data on vehicle position and movement.

Applications of Big Data

- Big Data for Financial Services

  Credit card companies, retail banks, private wealth management advisories, insurance firms, venture funds, and institutional investment banks all use big data for their financial services. The common problem among them all is the massive amounts of multi-structured data living in multiple disparate systems, which big data can solve. As such, big data is used in several ways, including:

1. Customer analytics
2. Compliance analytics

3. Fraud analytics
4. Operational analytics

- Big Data in Communications

  Gaining new subscribers, retaining customers, and expanding within current subscriber bases are top priorities for telecommunication service providers. The solutions to these challenges lie in the ability to combine and analyze the masses of customer-generated data and machine-generated data that is being created every day.

- Big Data for Retail

  Whether it's a brick-and-mortar company an online retailer, the answer to staying in the game and being competitive is understanding the customer better. This requires the ability to analyze all disparate data sources that companies deal with every day, including the weblogs, customer transaction data, social media, store-branded credit card data, and loyalty program data.

Data Science	Big Data
Data Science is an area.	Big Data is a technique to collect, maintain and process huge information.
It is about the collection, processing, analyzing, and utilizing of data in various operations. It is more conceptual.	It is about extracting vital and valuable information from a huge amount of data.
It is a field of study just like Computer Science, Applied Statistics, or Applied Mathematics.	It is a technique for tracking and discovering trends in complex data sets.

Data Science	Big Data
The goal is to build data-dominant products for a venture.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.
Tools mainly used in Data Science include SAS, R, Python, etc	Tools mostly used in Big Data include Hadoop, Spark, Flink, etc.
It is a superset of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics, and many more techniques.	It is a sub-set of Data Science as mining activities which is in a pipeline of Data science.
It is mainly used for scientific purposes.	It is mainly used for business purposes and customer satisfaction.
It broadly focuses on the science of the data.	It is more involved with the processes of handling voluminous data.

Getting Past the Hype(promote)

- Rachel's experience going from getting a PhD in statistics to working at Google is a great example to illustrate why we thought, in spite of the above-mentioned reasons to be doubtful, there might be some meat in the data science sandwich. In her words:
- It was clear to me pretty quickly that the stuff I was working on at Google was different than anything I had learned at school when I got my PhD in statistics. This is not to say that my degree was useless; far from it—what I'd learned in school provided a framework and way of thinking that I relied on daily, and much of the actual content provided a solid theoretical and practical foundation necessary to do my work.

Datafication

- Datafication is a buzzword ie, popular/important word
- Datafication is the transformation of social action into online quantified data, thus allowing for real-time tracking and predictive analysis. Simply said, it is about taking previously invisible process/activity and turning it into data, that can be monitored, tracked, analysed and optimised. Latest technologies we use have enabled lots of new ways of 'datify' our daily and basic activities.
- Summarizing, datafication is a technological trend turning many aspects of our lives into computerized data using processes to transform organizations into data-driven enterprises by converting this information into new forms of value.
- Datafication refers to the fact that daily interactions of living things can be rendered into a data format and put to social use.
- Organizations require data and extract knowledge and information to perform critical business processes. An organization also uses data for decision making, strategies and other key objectives.
- Datafication needs that in a modern data-oriented landscape, an organization's survival is depending on total control over the storage, extraction, manipulation and extraction of data and associated information.

- Example :

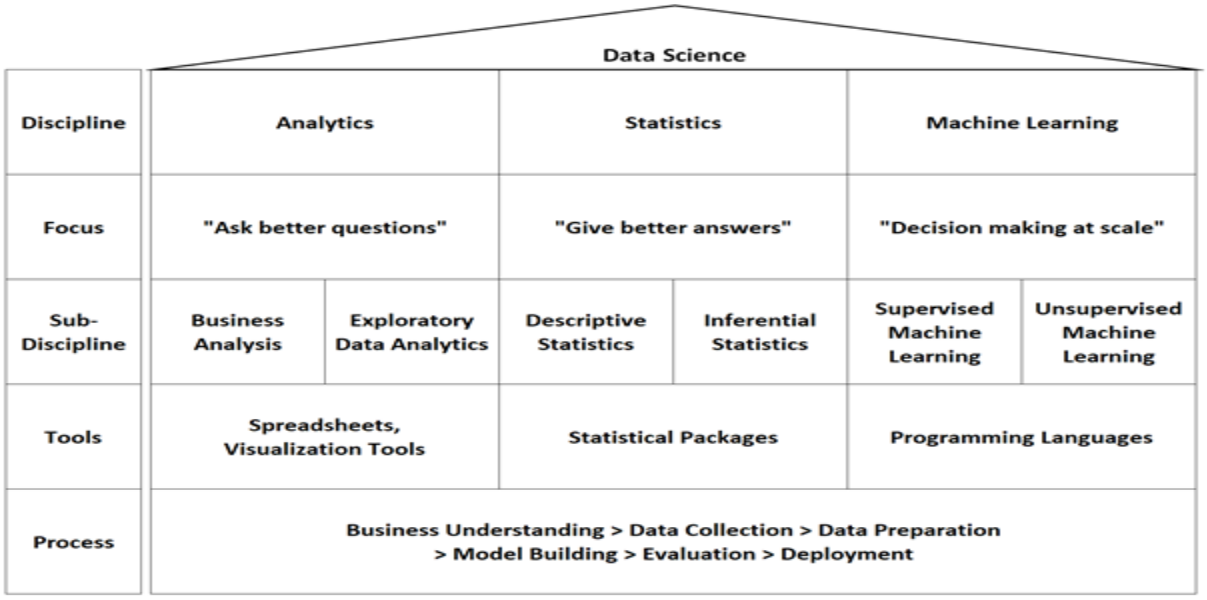
  For example, we create data every time we talk on the phone, SMS, tweet, email, use Facebook, watch a video, withdraw money from an ATM, use a credit card, or even walk past a security camera. The notion is different from digitization. In fact datafication is far broader than digitization. This astronomical amount of data has information about our identity and our behaviour.

- For example marketers are analysing Facebook and Twitter data to determine and predict sales. Companies spanning from all sectors and sizes have started to realize the big benefits of data and its analytics. They are beginning to improve their capabilities to collect and analyse data.

Current landscape of perspectives

The Data Science Landscape

Data science is part of the computer sciences . It comprises the disciplines of i) analytics, ii) statistics and iii) machine learning.



The Data Science Landscape

2.1. Analytics

- Analytics generates insights from data using simple presentation, manipulation, calculation or visualization of data. In the context of data science, it is also sometimes referred to as exploratory(trial) data analytics.

It often serves the purpose to familiarize oneself with the subject matter and to obtain some initial hints for further analysis.

- Business analysis is a professional discipline of identifying business needs and determining solutions to business problems.
- Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

2.2. Statistics

- Statistics is a branch of mathematics that is used for organization and interpretation of the numerical data. When it comes to the data organization, the kind of statistics we used are known as descriptive statistics.
- So descriptive statistics is basically used to describe the situation or the event, or whatever the property that we are measuring. For example, suppose we are discussing the marks obtained by the student in the examination, we might be interested in the average marks scored by the student, or the spread or division of the marks. So mean, median, standard deviation, percentile, etc. they all examples of descriptive statistics.

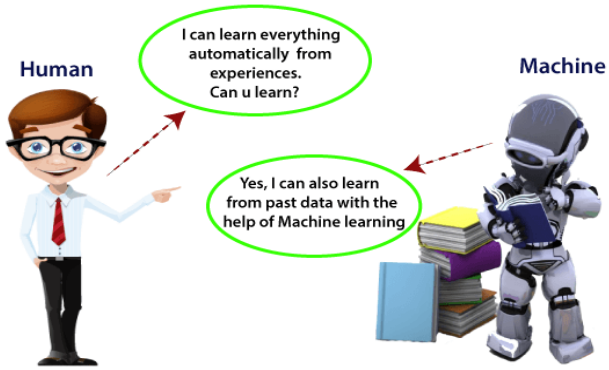
- **Descriptive statistics** summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the center of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).

- Measures of frequency distribution describe the occurrence of data within the data set (count)

- **Inferential statistics** can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. The goal of inferential statistics is to make generalizations about a population. In inferential statistics, a statistic is taken from the sample data (e.g., the sample mean) that used to make inferences about the population parameter (e.g., the population mean).

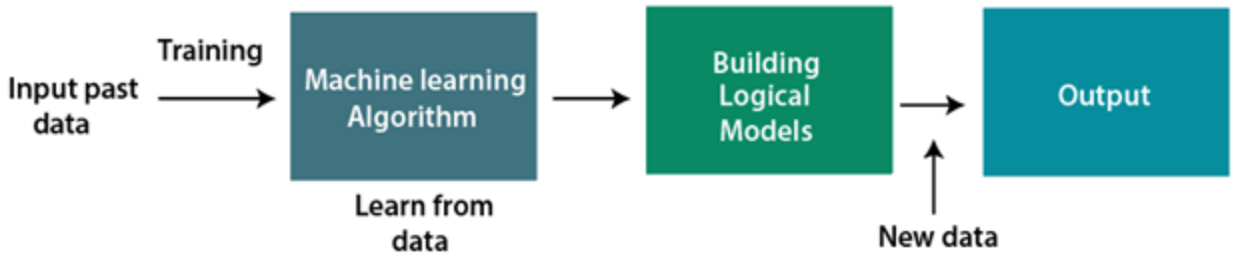
- ✓ **Artificial Intelligence** is composed of two words **Artificial** and **Intelligence**, where Artificial defines *"man-made,"* and intelligence defines *"thinking power"*, hence AI means *"a man-made thinking power."*
- ✓ "It is a branch of computer science by which we can create intelligent machines which can behave like a human, think like humans, and able to make decisions."
- ✓ Intelligence, as we know, is the ability to acquire and apply knowledge. Knowledge is the information acquired through experience. Experience is the knowledge gained through exposure(training). Summing the terms up, we get **artificial intelligence** as the "copy of something natural(i.e., human beings) 'WHO' is capable of acquiring and applying the information it has gained through exposure."

Machine Learning



- ✓ Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own.
- ✓ Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

- ✓ A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.
- ✓ Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem.



Supervised Machine Learning

- ✓ Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.
- ✓ **Supervised learning** is when the model is getting trained on a labelled dataset. A **labelled** dataset is one that has both input and output parameters. In this type of learning both training and validation, datasets are labelled
- ✓ In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

E.G: Image classification, Fraud Detection, spam filtering, etc.

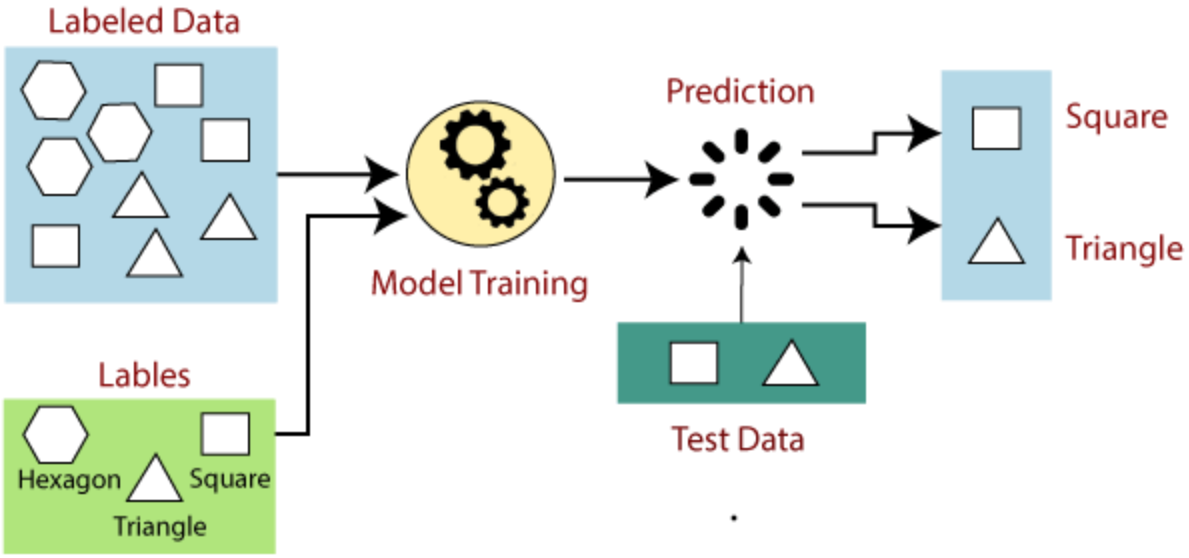


Fig. Example for Supervised learning

Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.

- ✓ If the given shape has four sides, and all the sides are equal, then it will be labelled as a **Square**.
- ✓ If the given shape has three sides, then it will be labelled as a **triangle.**
- ✓ If the given shape has six equal sides then it will be labelled as **hexagon**.
- ✓ Now, after training, we test our model using the test set, and the task of the model is to identify the **shape**.
- ✓ The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

Both the above figures have labelled data set as follows:

- Figure A: It is a dataset of a shopping store that is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/ her gender, age, and salary.
  Input: Gender, Age, Salary
  Output: Purchased i.e. 0 or 1; 1 means yes the customer will purchase and 0 means that the customer won't purchase it.
- Figure B: It is a Meteorological dataset that serves the purpose of predicting wind speed based on different parameters.
  Input: Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction
  Output: Wind Speed

Training the system: While training the model, data is usually split in the ratio of 80:20 i.e. 80% as training data and the rest as testing data. In training data, we feed input as well as output for 80% of data. The model learns from training data only. We use different machine learning algorithms to build our model. Learning means that the model will build some logic of its own. Once the model is ready then it is good to be tested. At the time of testing, the input is fed from the remaining 20% of data that the model has never seen before, the model will predict some value and we will compare it with the actual output and calculate the accuracy.

Unsupervised Learning

- *Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision*

Example: Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images.

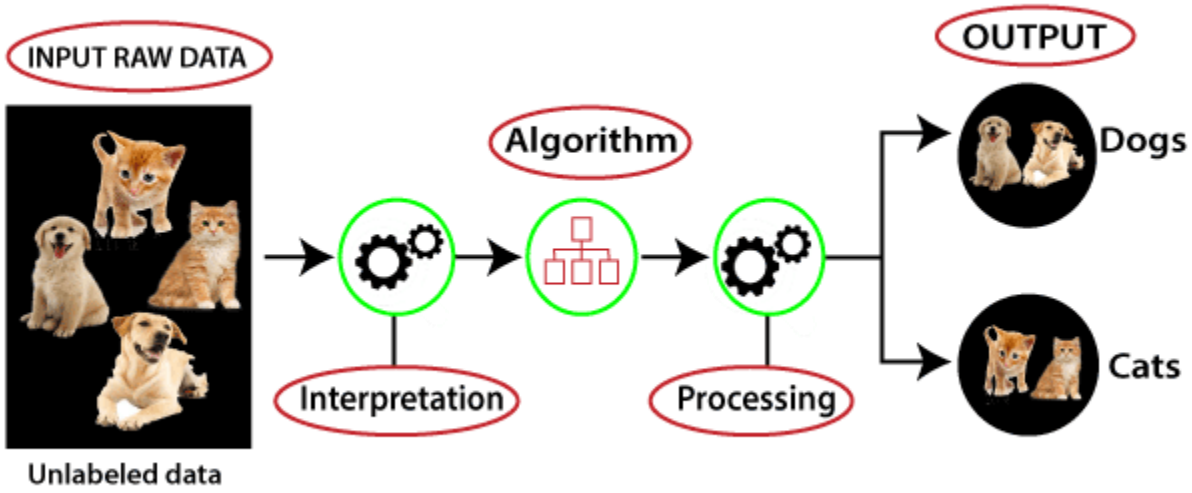


Fig. Example for Unsupervised learning

- Here, we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.

- ❖ Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

Statistical Inference

- ❖ Statistics is a branch of Mathematics, that deals with the collection, analysis, interpretation, and the presentation of the numerical data. In other words, it is defined as the collection of quantitative data.
- ❖ The main purpose of Statistics is to make an accurate conclusion using a limited sample about a greater population.

Types of Statistics

Statistics can be classified into two different categories. The two different types of Statistics are:

- Descriptive Statistics
- Inferential Statistics
- **Descriptive statistics** summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the center of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).
- Measures of frequency distribution describe the occurrence of data within the data set (count)

- **Inferential statistics** can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. The goal of inferential statistics is to make generalizations about a population. In inferential statistics, a statistic is taken from the

sample data (e.g., the sample mean) that used to make inferences about the population parameter (e.g., the population mean).

- In Statistics, **descriptive statistics** describe the data, whereas **inferential statistics** help you make predictions from the data. In inferential statistics, the data are taken from the sample and allows you to generalize the population.
- In general, inference means "guess", which means making inference about something. So, statistical inference means, making inference about the population. To take a conclusion about the population, it uses various statistical analysis techniques.

Populations and samples

Population:

- A complete collection of the objects or measurements is called the population or else everything in the group we want to learn about will be termed as population or else In statistics population is the entire set of items from which data is drawn in the statistical study.
- It can be a group of individuals or a set of items.

Population is the entire group you want to draw conclusions about.

- ✓ The population is usually denoted with **N**
- ✓ The number of citizens living in the State of Rajasthan represents a population of the state
- ✓ the number of planets in the entire universe represents the planet population of the entire universe
- ✓ The types of candies and chocolates are made in India.
- ✓ population mean is usually denoted by the Greek letter μ

μ (population mean) = $\sum_{i=1}^{N}(xi)$ / N(total population)

For example: let us assume that there are 5 employees in my company, so 5 people is a complete set hence it will represent the population of my company

If I want to find the average age of my company then I will simply add their ages and divide it by N which is the number of population

ages = {23,45,12,34,22}

$$\mu = \sum_{i=1}^{5}(xi)/5$$

$$= (23 + 45 + 12 + 34 + 22) / 5$$

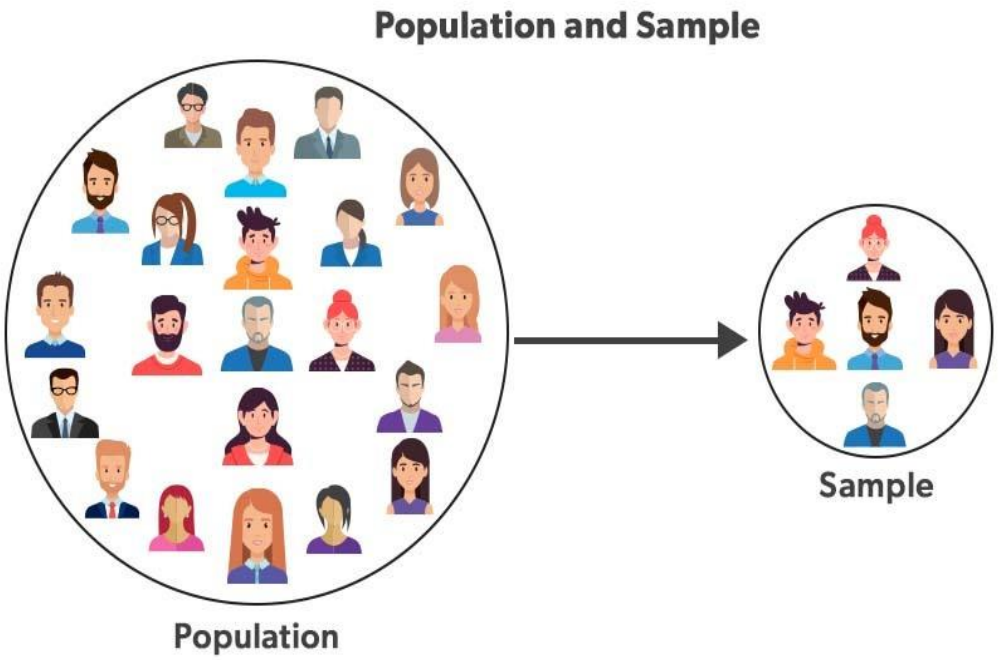
the results say that the average age of my company is 27.2 years

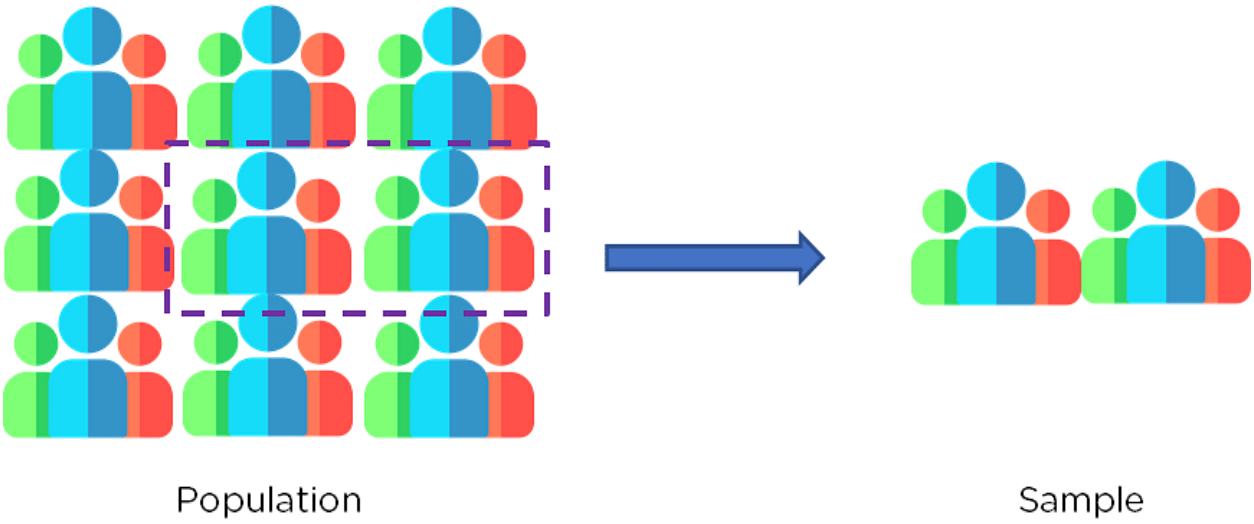
so this is what we call the **population mean**

Sample:

- A sample represents a group of the interest of the population which we will use to represent the data. The sample is an unbiased(balanced) subset of the population in which we represent the whole data.
- A sample is a group of the elements actually participating in the survey or study.
- A sample is the representation of the manageable size.
- samples are collected and stats are calculated from the sample so one can make interferences or extrapolations from the sample.

- This process of collecting information from the sample is called **sampling.**





- ✓ *The sample is denoted by the* ***n***
- ✓ 500 people from a total population of the Rajasthan state will be considered as a sample

- ✓ 143 total chess players from all total number of chess players will be considered as a sample
- ✓ Sample mean is denoted by x –

$$x\ (\textbf{\textit{sample mean}}) = \sum_{i=1}^{n}(xi)/n\ (\text{total sample})$$

Example: Let us assume the population of India is 10 million, and recent elections were conducted in India between two parties 'party A ' and 'party B' now researchers want to find which party is winning so here we will create a group of few people lets say 10,000 from different regions and age groups so that sample is not biased. Then ask them who they voted we can get the exit poll. This is the thing which most of our media do during the elections, and show stats such as there 55% chances of party A winning the elections.

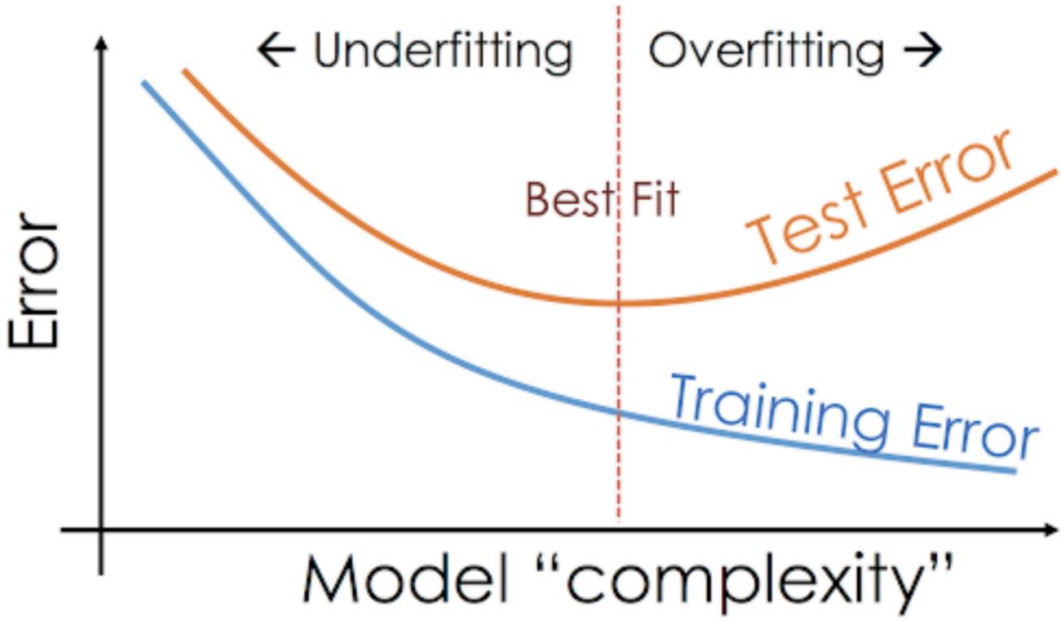
Samples are used when :

- The population is too large to collect data.
- The data collected is not reliable.
- The population is hypothetical(Proposed) and is unlimited in size.

Fitting a model – Over fitting

- Overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data.
- When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose.
- When machine learning algorithms are constructed, they leverage(control) a sample dataset to train the model.
- However, when the model trains for too long on sample data or when the model is too complex, it can start to learn the "noise," or irrelevant information, within the dataset.
- When the model memorizes the noise and fits too closely to the training set, the model becomes "overfitted," and it is unable to generalize well to new data.

- If a model cannot generalize well to new data, then it will not be able to perform the classification or prediction tasks that it was intended for.
- If the training data has a low error rate and the test data has a high error rate, it signals overfitting.



- **How to avoid overfitting**

- **Early stopping:** This method seeks to pause training before the model starts learning the noise within the model.
- **Train with more data:** Expanding the training set to include more data can increase the accuracy of the model. this is a more effective method when clean, relevant data is injected into the model
- **Data augmentation:** While it is better to inject clean, relevant data into your training data, sometimes noisy data is added to make a model more stable.
- Which is used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from

existing data. It acts as a regularizer and helps reduce overfitting when training a machine learning model.

- However, this method should be done sparingly(in a restricted or in small quantities).
- **Regularization:** If overfitting occurs when a model is too complex, it makes sense for us to reduce the number of features.
- But what if we don't know which inputs to eliminate during the feature selection process. So certain If we don't know which features to remove from our model, **regularization methods** can be particularly helpful
- E.G L1 regularization(Lasso regularization)

Statistical modeling

- *Statistical Modelling is a process of using data to construct a mathematical or algorithmic device to measure the probability of some observation*
- ***Training:*** *Using an set of observations to learn parameters of a model or construct the decision making process.*
- ***Evaluation:***
- *Determining the probability of a new observation*

  - *Statistical Modelling is simply the method of implementing statistical analysis to a dataset*

- *where a Statistical Model is a mathematical representation of observed data."*
- Statistical modeling refers to the data science process of applying statistical analysis to datasets.

Statistical Modeling Techniques

- The first step in developing a statistical model is gathering data, which may be sourced from spreadsheets, databases, data lakes, or the cloud.

- The most common statistical modeling methods for analyzing this data are categorized as either supervised learning or unsupervised learning.
- Some **popular statistical model examples** include logistic regression, time-series, clustering, and decision trees.

Supervised learning techniques include **regression models and classification models**:

Ex:

- Prediction of rain using temperature and other factors
- Determining Market trends

Terminologies Related to the Regression Analysis:

- **Dependent Variable(target variable):** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **EX: A test score** could be a dependent variable because it could change depending on several factors such as how much you studied, how much sleep you got the night before you took the test, or even how hungry you were when you took it.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **EX:** It is a variable that stands alone and isn't changed by the other variables you are trying to measure. For example, someone's age might be an independent variable. Other factors (such as what they eat, how much they go to school, how much television they watch) aren't going to change a person's age.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

- ✓ Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of **Market Trends**, prediction of House prices, etc.
- ✓ The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).

Example: Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

Types of Regression Algorithm:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

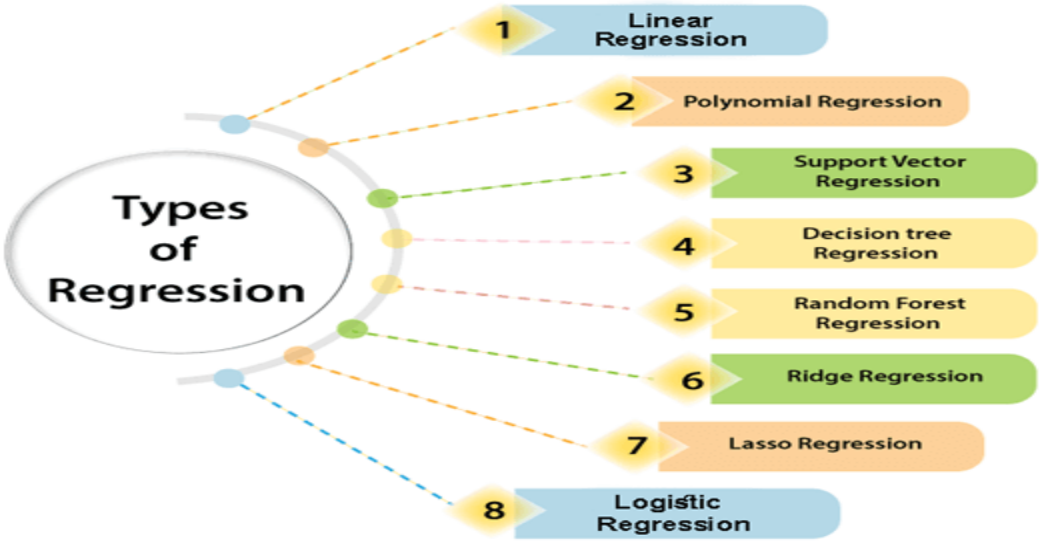
Purpose of Regression Analysis

- ❖ Regression analysis helps in the prediction of a continuous variable.
- ❖ There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately.

- So for such case we need Regression analysis which is a statistical method and used in machine learning and data science.
- Regression estimates the relationship between the target and the independent variable.

Reasons

- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.



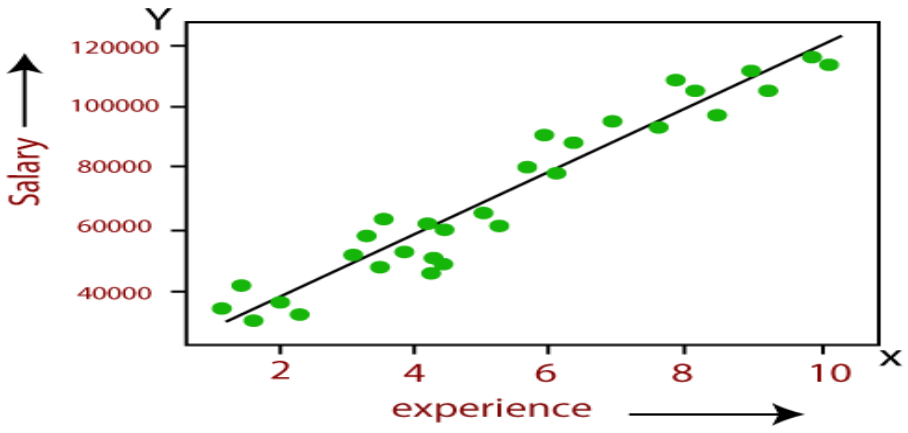
Linear Regression:

- Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of Market Trends, prediction of House prices, etc.
- The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).
- Example: Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

- Linear regression is a statistical regression method which is used for predictive analysis. Linear regression is a machine learning concept that is

used to build or train the models (mathematical models or equations) for solving **supervised learning problems** related to predicting **continuous numerical value**

- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression analysis is used to predict the value of a variable based on the value of another variable.
- The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



- Below is the mathematical equation for Linear regression:

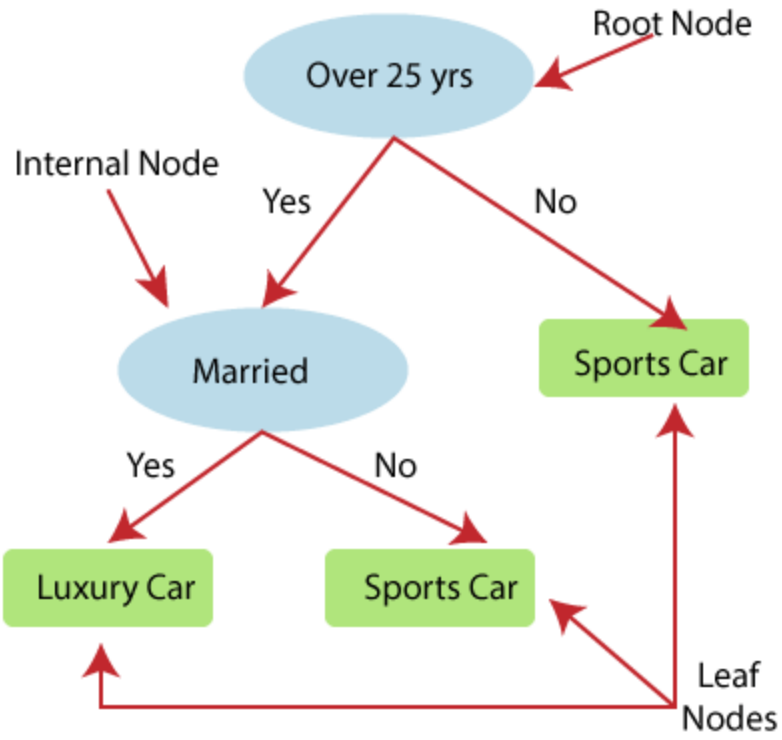
1. Y= aX+b

- **Here, Y = dependent variables (target variables), X= Independent variables (predictor variables), a and b are the linear coefficients**
- Some popular applications of linear regression are:

- **Analyzing trends and sales estimates**
- **Salary forecasting**
- **Real estate prediction**

Decision Tree Regression:

- Decision Tree is a supervised learning algorithm which can be used for solving both classification and regression problems.
- It can solve problems for both categorical and numerical data
- Decision Tree regression builds a tree-like structure in which each internal node represents the "test" for an attribute, each branch represent the result of the test, and each leaf node represents the final decision or result.
- A decision tree is constructed starting from the root node/parent node (dataset), which splits into left and right child nodes (subsets of dataset). These child nodes are further divided into their children node, and themselves become the parent node of those nodes. Consider the below image:

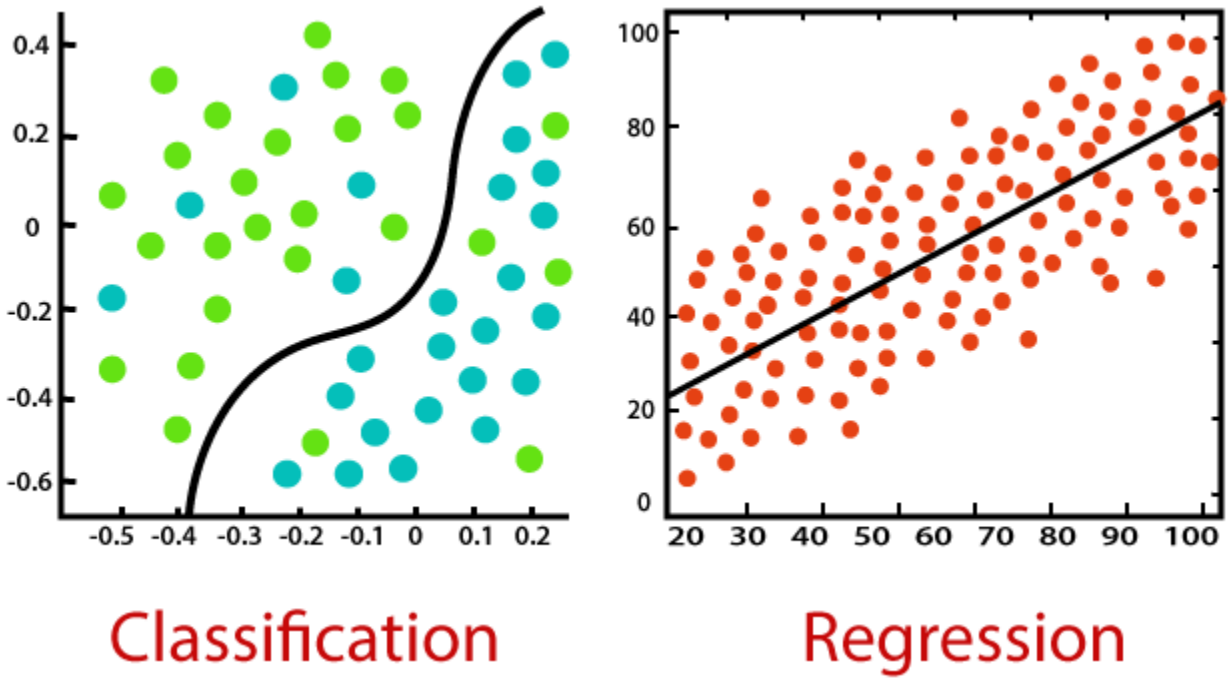


- Above image showing the example of Decision Tee regression, here, the model is trying to predict the choice of a person between Sports cars or Luxury car.

Notes additional :

- Regression and Classification algorithms are Supervised Learning algorithms. Both the algorithms are used for prediction in Machine learning and work with the labeled datasets. But the difference between both is how they are used for different machine learning problems.
- The main difference between Regression and Classification algorithms that Regression algorithms are used to **predict the continuous** values such as price, salary, age, etc. and Classification algorithms are used to **predict/Classify the discrete values** such as Male or Female, True or False, Spam or Not Spam, etc.

Consider the below diagram:



Classification:

- Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters. In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes.

Example: The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.

Types of ML Classification Algorithms:

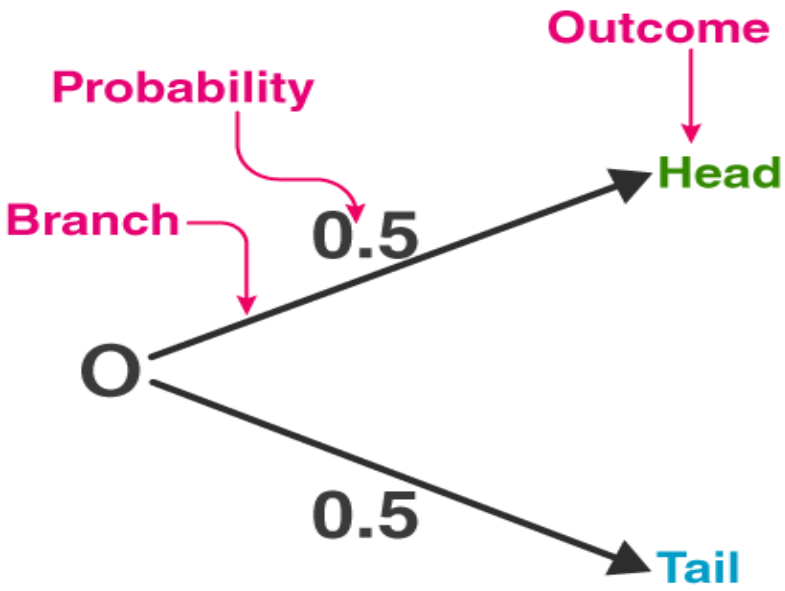
Classification Algorithms can be further divided into the following types:

- ✓ Logistic Regression
- ✓ K-Nearest Neighbours
- ✓ Support Vector Machines

- ✓ Kernel SVM
- ✓ Naïve Bayes
- ✓ Decision Tree Classification
- ✓ Random Forest Classification

Probability

- The word 'Probability' means the chance of occurring of a particular event.
- Probability denotes the possibility of something happening.
- It is a mathematical concept that predicts how likely events are to occur. The probability values are expressed between 0 and 1.
- The definition of probability is the degree to which something is likely to occur.



Important Terms related to Probability:

1. Trial and Event: The **performance of an experiment** is called a trial, and the set of its outcomes is termed an event.

Example: Tossing a coin and getting head is a trial. Then the event is {HT, TH, HH}

2. Random Experiment: It is an experiment in which all the possible outcomes of the experiment are known in advance. But the exact outcomes of any specific performance are not known in advance.

Example:

1. Tossing a Coin
2. Rolling a dice
3. Drawing a card from a pack of 52 cards.
4. Drawing a ball from a bag.

3. Outcome: The result of a random experiment is called an Outcome.

Example: 1. Tossing a coin is an experiment and getting head is called an outcome.

2. Rolling a dice and getting 6 is an outcome.

4. Sample Space: The set of all possible outcomes of an experiment is called sample space and is denoted by S.

Example: When a dice is thrown, sample space is S = {1, 2, 3, 4, 5, 6}

It consists of six outcomes 1, 2, 3, 4, 5, 6

Probability distribution

- Probability distribution is a function that is used to give the **probability of all the possible values that a random variable can take**.
- A **probability distribution** is a mathematical function that describes the **probability of different possible values of a variable**.
- Probability distributions are often depicted using **graphs or probability tables**.
- probability distribution gives the **possibility of each outcome of a random experiment or event.**
- It provides the probabilities of **different possible occurrences**.
- A probability distribution is a statistical function that **describes all the possible values and probabilities for a random variable within a given range.**
- This range will be bound by the minimum and maximum possible values, but where the possible value would be plotted on the probability **distribution will be determined by a number of factors**

Types of Probability Distribution

The probability distribution is divided into two parts:

- Discrete Probability Distributions
- Continuous Probability Distributions

1.Discrete Probability Distributions

- A discrete probability distribution can be defined as a probability distribution giving the probability that a discrete random variable will have a specified value.
- Such a distribution will represent data that has a finite countable number of outcomes.

Discrete Probability Distribution Example

- **EX:1** A discrete distribution has a range of values that are countable. For example, the numbers on birthday cards have a possible range from 0 to 122 (122 is the age of Jeanne the oldest person who ever lived)
- **EX:2** Suppose a fair dice is rolled and the discrete probability distribution has to be created. The possible outcomes are {1, 2, 3, 4, 5, 6}.
- Thus, the total number of outcomes will be 6. All numbers have a fair chance of turning up.
- This means that the probability of getting any one number is 1 / 6. Using this data the discrete probability distribution table for a dice roll can be given as follows:

x	1	2	3	4	5	6
P(X = x)	1 / 6	1 / 6	1 / 6	1 / 6	1 / 6	1 / 6

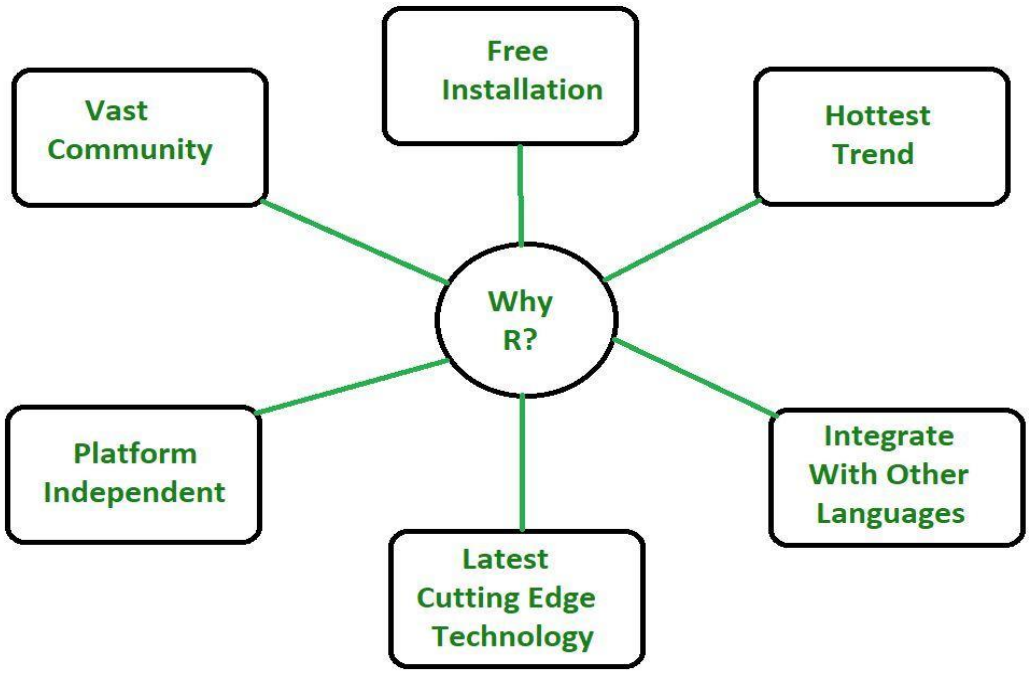
2. Continuous Probability Distributions

- A continuous distribution has a range of values that are infinite, and therefore uncountable.
- For example, time is infinite: you could count from 0 seconds to a billion seconds…a trillion seconds…and so on,

Basics of R: Introduction

- ✓ R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface. R is available across widely used platforms like Windows, Linux, and macOS. Also, the R programming language is the latest cutting-edge tool.
- ✓ It was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R programming language is an implementation of the S programming language. It also combines with lexical scoping semantics inspired by Scheme. Moreover, the project conceives in 1992, with an initial version released in 1995 and a stable beta version in 2000.

Why R Programming Language?



- R programming is used as a leading tool for machine learning, statistics, and data analysis. Objects, functions, and packages can easily be created by R.
- It's a platform-independent language. This means it can be applied to all operating system.
- It's an open-source free language. That means anyone can install it in any organization without purchasing a license.
- R programming language is not only a statistic package but also allows us to integrate with other languages (C, C++). Thus, you can easily interact with many data sources and statistical packages.
- The R programming language has a vast community of users and it's growing day by day.
- R is currently one of the most requested programming languages in the Data Science job market that makes it the hottest trend nowadays.

Features of R Programming Language

Statistical Features of R:

- **Basic Statistics:** The most common basic statistics terms are the mean, mode, and median. These are all known as "Measures of Central Tendency." So using the R language we can measure central tendency very easily.

- **Static graphics:** R is rich with facilities for creating and developing interesting static graphics. R contains functionality for many plot types including graphic maps, mosaic plots, biplots, and the list goes on.
- **Probability distributions:** Probability distributions play a vital role in statistics and by using R we can easily handle various types of probability distribution such as Binomial Distribution, Normal Distribution, Chi-squared Distribution and many more.
- **Data analysis:** It provides a large, coherent and integrated collection of tools for data analysis.

Programming Features of R:

- **R Packages:** One of the major features of R is it has a wide availability of libraries. R has CRAN(Comprehensive R Archive Network), which is a repository holding more than 10, 0000 packages.
- **Distributed Computing:** Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Two new packages ddR and multidplyr used for distributed programming in R were released in November 2015.
    - **Programming in R:**
    - Since R is much similar to other widely used languages syntactically, it is easier to code and learn in R. Programs can be written in R in any of the widely used IDE like **R Studio, Rattle, Tinn-R**, etc. After writing the program save the file with the extension .r. To run the program use the following command on the command line:
    - `R file_name.r`

Advantages of R:

- R is the most comprehensive statistical analysis package. As new technology and concepts often appear first in R.
- As R programming language is an open source. Thus, you can run R anywhere and at any time.
- R programming language is suitable for GNU/Linux and Windows operating system.
- R programming is cross-platform which runs on any operating system.

- In R, everyone is welcome to provide new packages, bug fixes, and code enhancements.

Disadvantages of R:

- In the R programming language, the standard of some packages is less than perfect.
- Although, R commands give little pressure to memory management. So R programming language may consume all available memory.
- In R basically, nobody to complain if something doesn't work.
- R programming language is much slower than other programming languages such as Python and MATLAB.

Applications of R:

- We use R for Data Science. It gives us a broad variety of libraries related to statistics. It also provides the environment for statistical computing and design.
- R is used by many quantitative analysts as its programming tool. Thus, it helps in data importing and cleaning.
- R is the most prevalent language. So many data analysts and research programmers use it. Hence, it is used as a fundamental tool for finance.
- Tech giants like Google, Facebook, bing, Twitter, Accenture, Wipro and many more using R nowadays.

Note: **"R is an interpreted computer programming language which was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand."** The ***R Development Core Team*** currently develops R. It is also a software environment used to analyze **statistical information**, **graphical representation**, **reporting**, and **data modeling**

R- Environment Setup

- **R programming** is a very popular language and to work on that we have to install two things, i.e., R and RStudio. R and RStudio works together to create a project on R.
- Installing R to the local computer is very easy. First, we must know which operating system we are using so that we can download it accordingly.

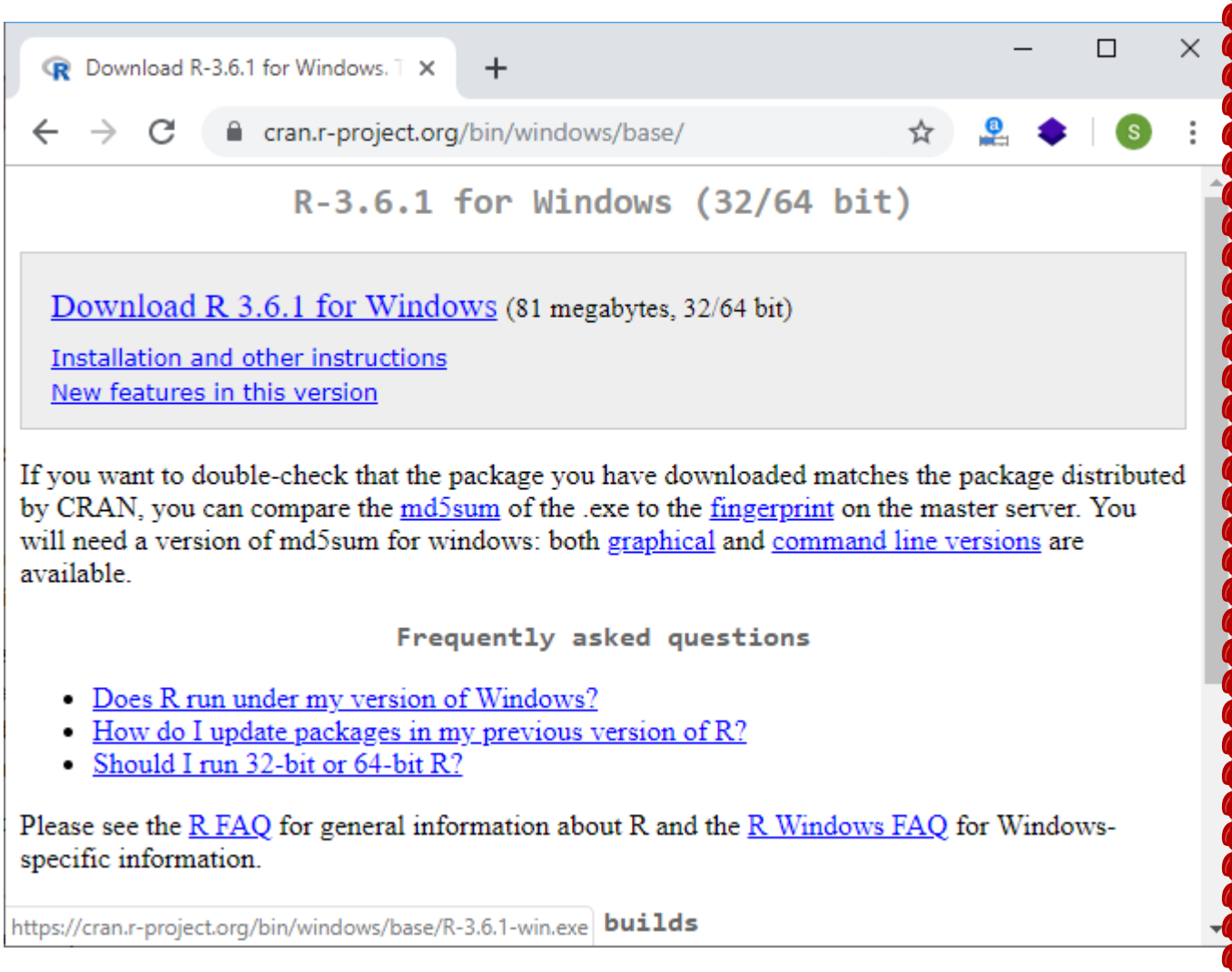
- The official site https://cloud.r-project.org provides binary files for major operating systems including Windows, Linux, and Mac OS. In some Linux distributions, R is installed by default, which we can verify from the console by entering R.
- To install R, either we can get it from the site https://cloud.r-project.org or can use commands from the terminal.

Install R in Windows

There are following steps used to install the R in Windows:

Step 1:

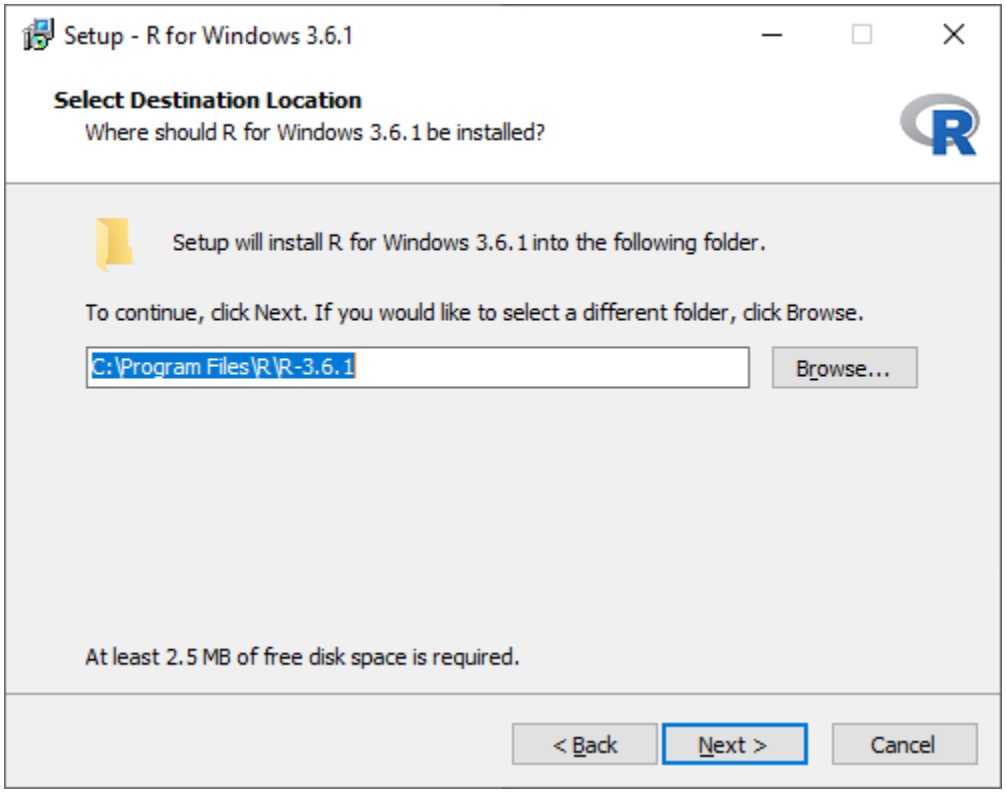
First, we have to download the R setup from https://cloud.r-project.org/bin/windows/base/.



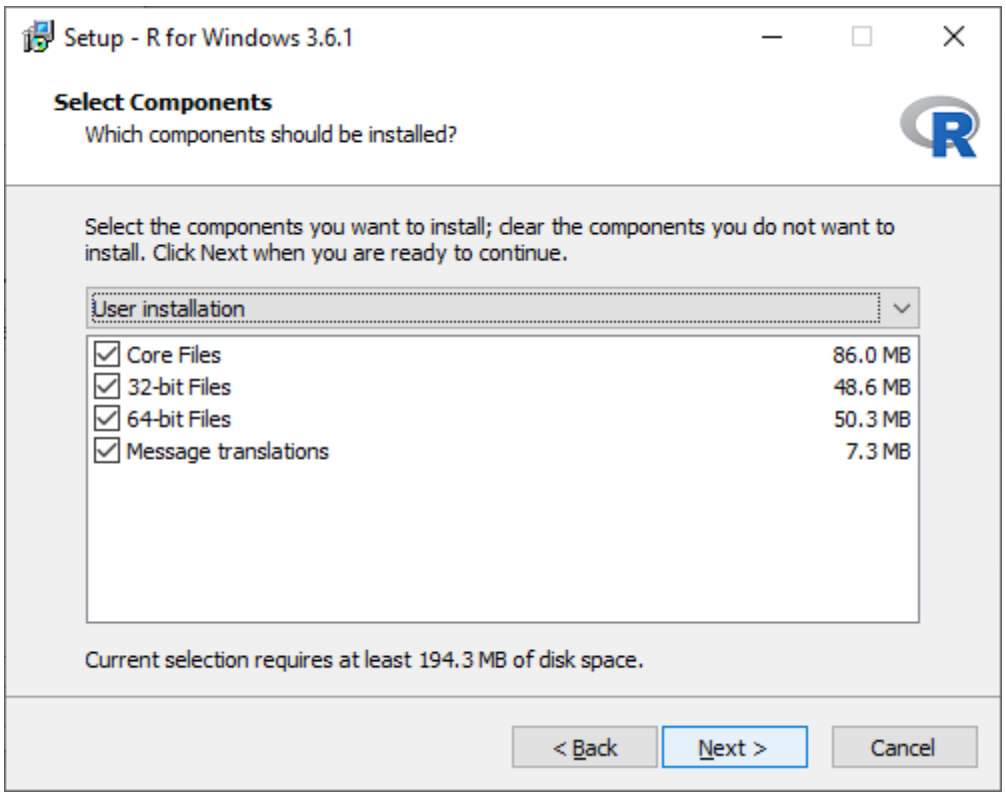
Step 2:

When we click on **Download R 3.6.1 for windows**, our downloading will be started of R setup. Once the downloading is finished, we have to run the setup of R in the following way:

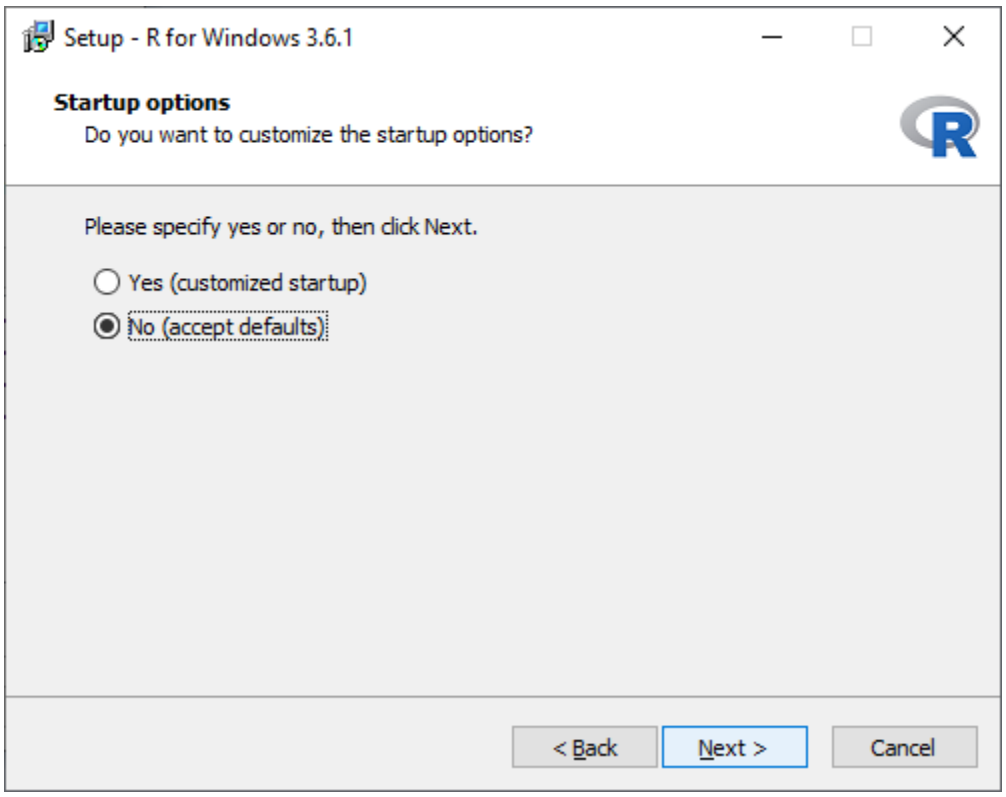
1) Select the path where we want to download the R and proceed to Next.



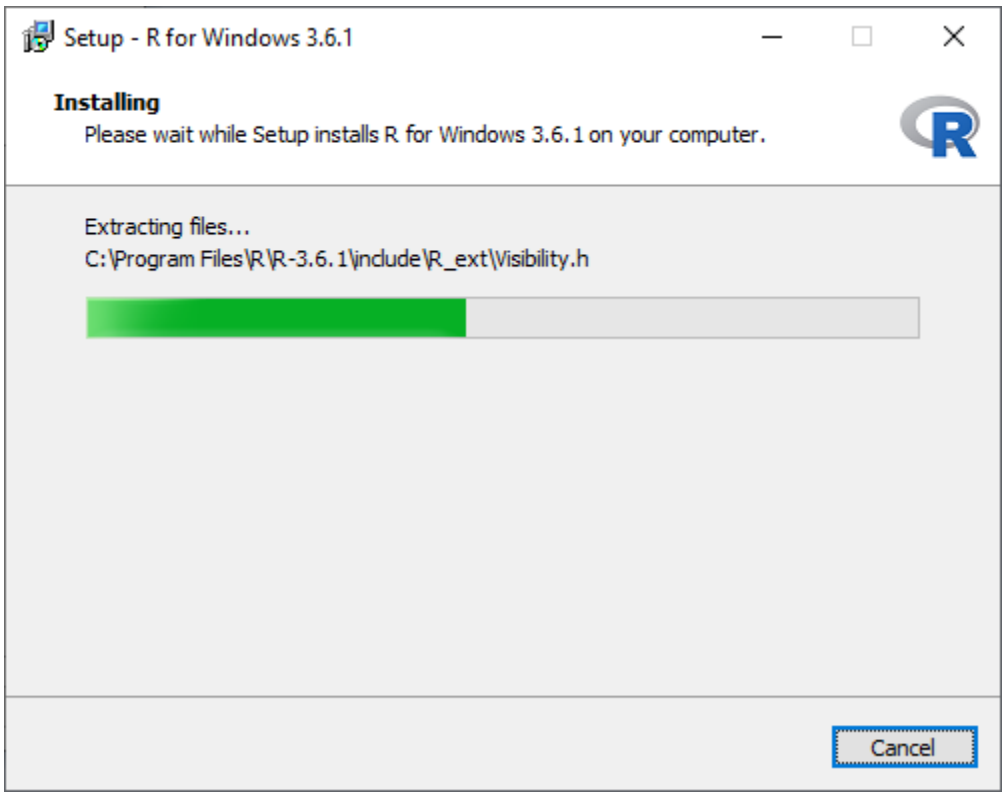
2) Select all components which we want to install, and then we will proceed to **Next**.



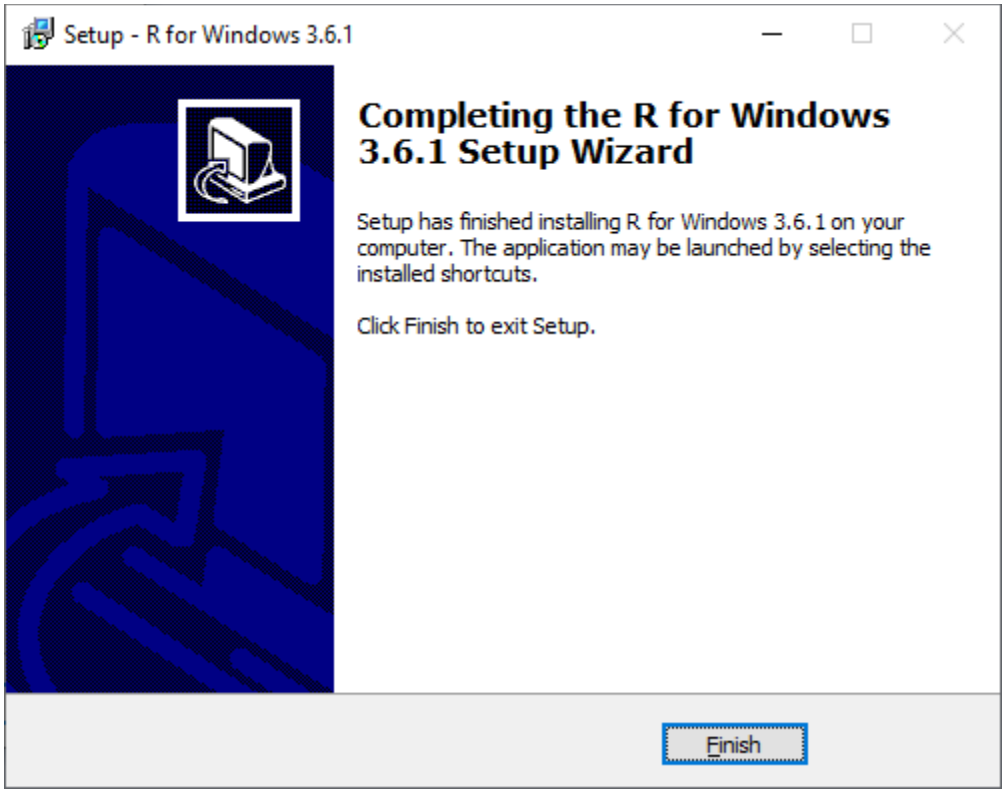
3) In the next step, we have to select either customized startup or accept the default, and then we proceed to **Next**.



4) When we proceed to next, our installation of R in our system will get started:



5) In the last, we will click on finish to successfully install R in our system.

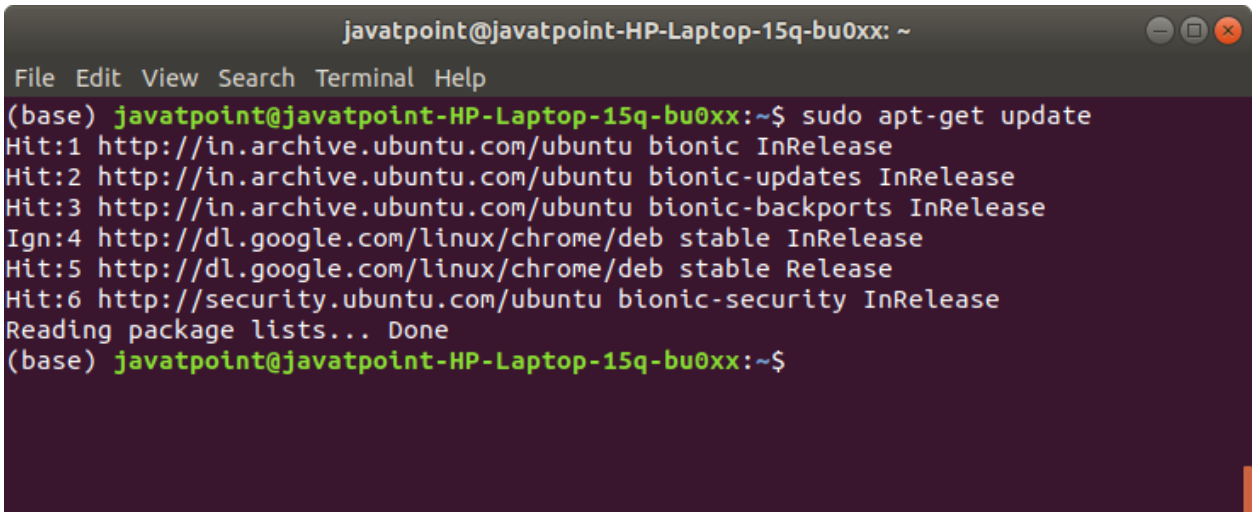


Install R in Linux

There are only three steps to install R in Linux

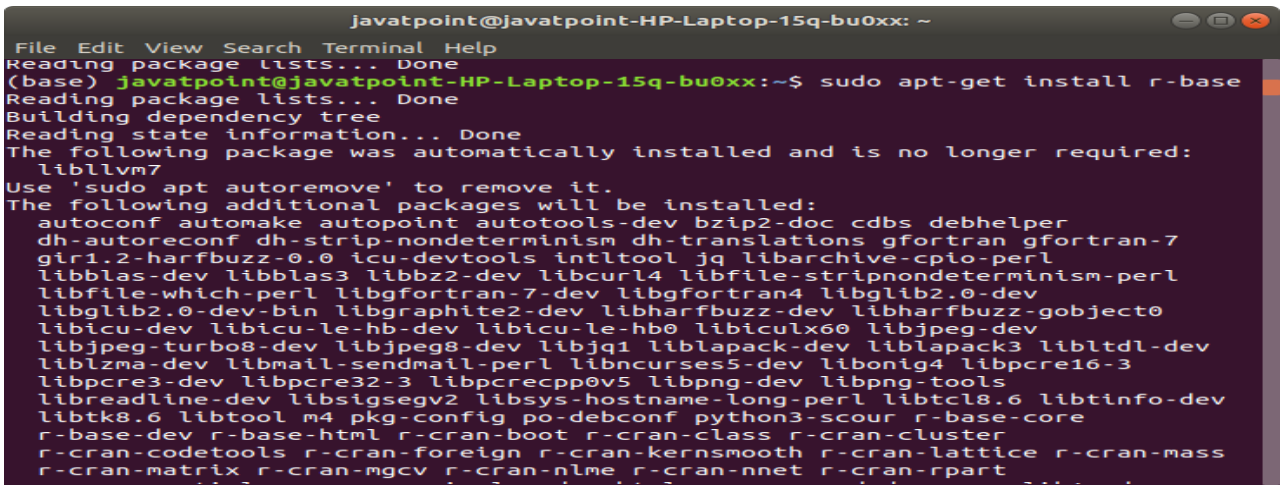
Step 1:

In the first step, we have to update all the required files in our system using **sudo apt-get update** command as:



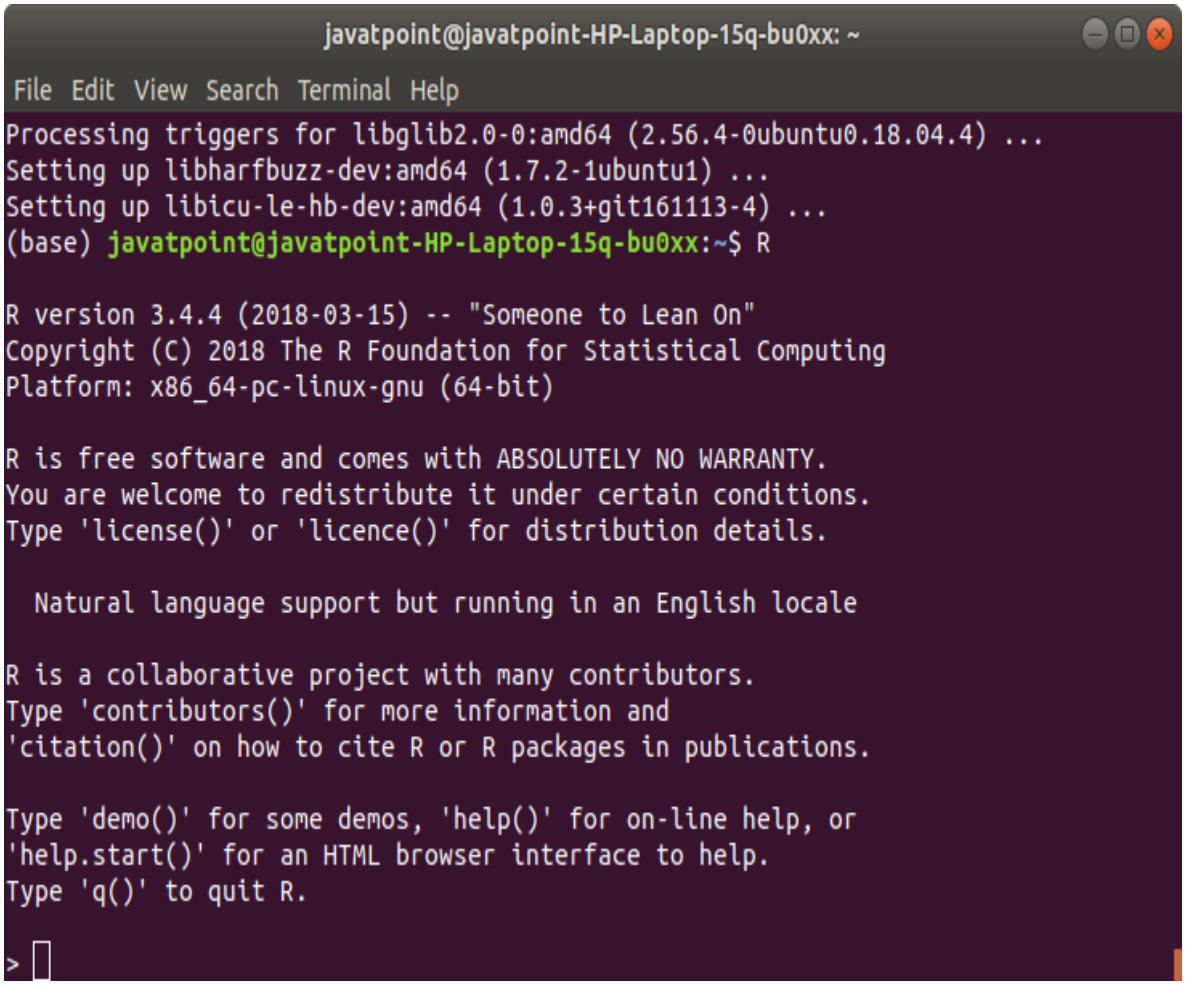
Step 2:

In the second step, we will install R file in our system with the help of **sudo apt-get install r-base** as:



Step 3:

In the last step, we type R and press enter to work on R editor.



RStudio IDE

RStudio is an integrated development environment which allows us to interact with R more readily. RStudio is similar to the standard RGui, but it is considered more user-friendly. This IDE has various drop-down menus, Windows with multiple tabs, and so many customization processes. The first time when we open RStudio, we will see three Windows. The fourth Window will be hidden by default. We can open this hidden Window by clicking the **File** drop-down menu, then **New File** and then **R Script**.

RStudio Windows/Tabs	Location	Description
Console Window	Lower-left	The location where commands are entered and output is printed.
Source Tabs	Upper-left	Built-in test editor
Environment Tab	Upper-left	An interactive list of loaded R objects.
History Tab	Upper-left	List of keystrokes entered into the console.
Files Tab	Lower-right	File explorer to navigate C drive folders.
Plots Tab	Lower-right	Output location for plots.
Packages Tab	Lower-right	List of installed packages.
Help Tab	Lower-right	Output location for help commands and help search Window.
Viewer Tab	Lower-right	Advanced tab for local web content.

Installation of RStudio

RStudio Desktop is available for both Windows and Linux. The open-source RStudio Desktop installation is very simple to install on both operating systems. The licensed version of RStudio has some more features than open-source.

Before installing RStudio, let's see what are the additional features in the license version of RStudio.

Factor	Open-Source	Commercial License
Overview	1) Access RStudio locally	All of the features of open-source are include with 1) There is a commercial license for organizations which are not able to use AGPL software. 2) It provides access to priority support.
	2) Code completion, syntax highlighting, and smart indentation	
	3) Can execute R code directly from the source editor	
	4) Quickly jump to function definitions.	
	5) Easily manage multiple working directories using projects.	
	6) Integrated R help and documentation.	
	7) Provide interactive debugger to diagnose and fix errors quickly.	
	8) Extensive package deployment tools.	
Support	It supports for community forums only.	1) It supports priority email. 2) It supports for an 8-hour response during business hour.
License	AGPL v3	RStudio License Agreement

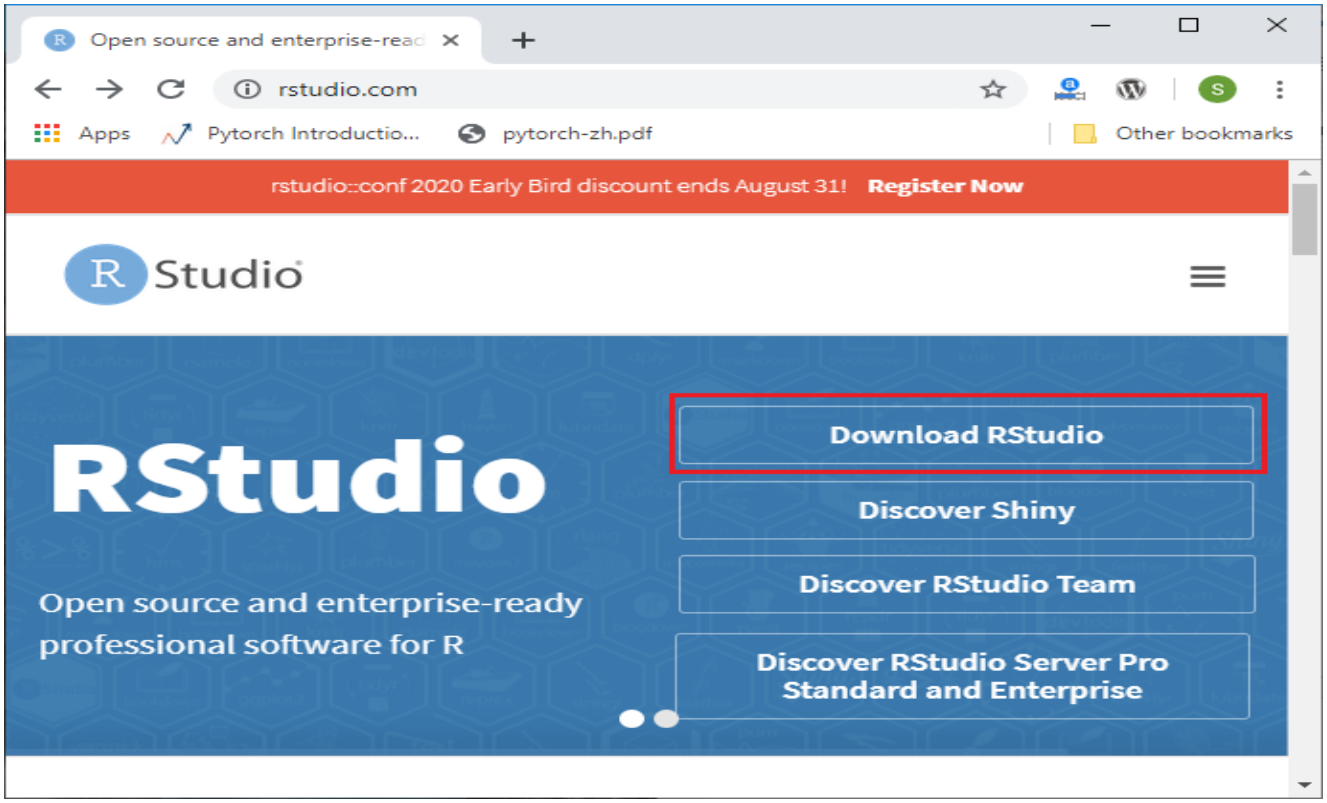
Pricing	Free	$995/year

Installation on Windows/Linux

- ✓ On Windows and Linux, it is quite simple to install RStudio. The process of installing RStudio in both the OS is the same. There are the following steps to install RStudio in our Windows/Linux:

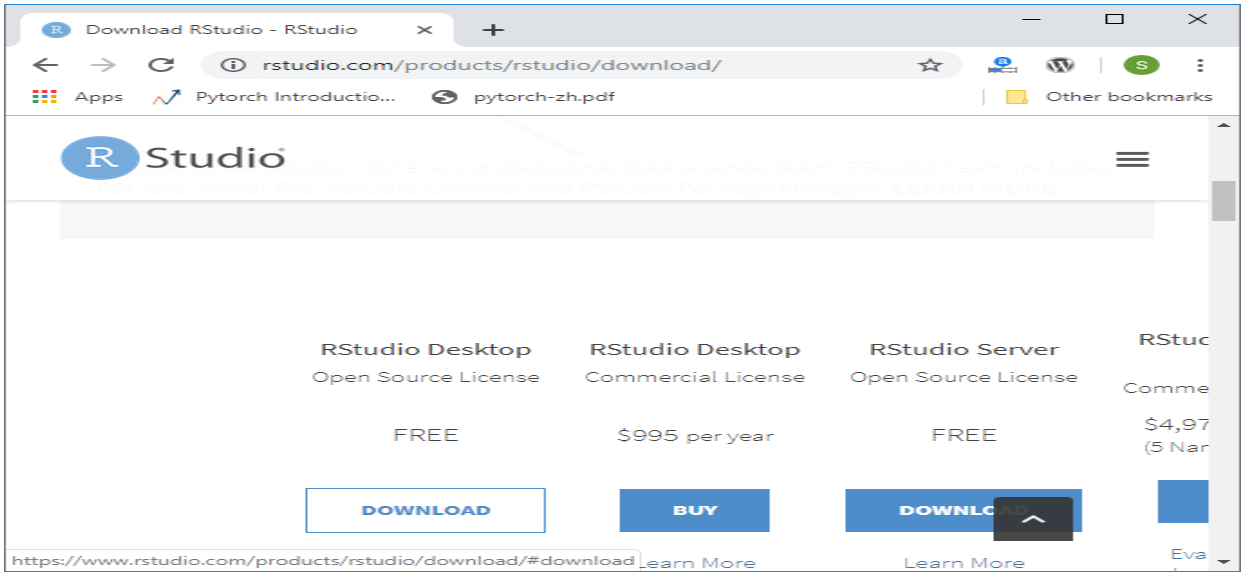
Step 1:

In the first step, we visit the RStudio official site and click on **Download RStudio**.



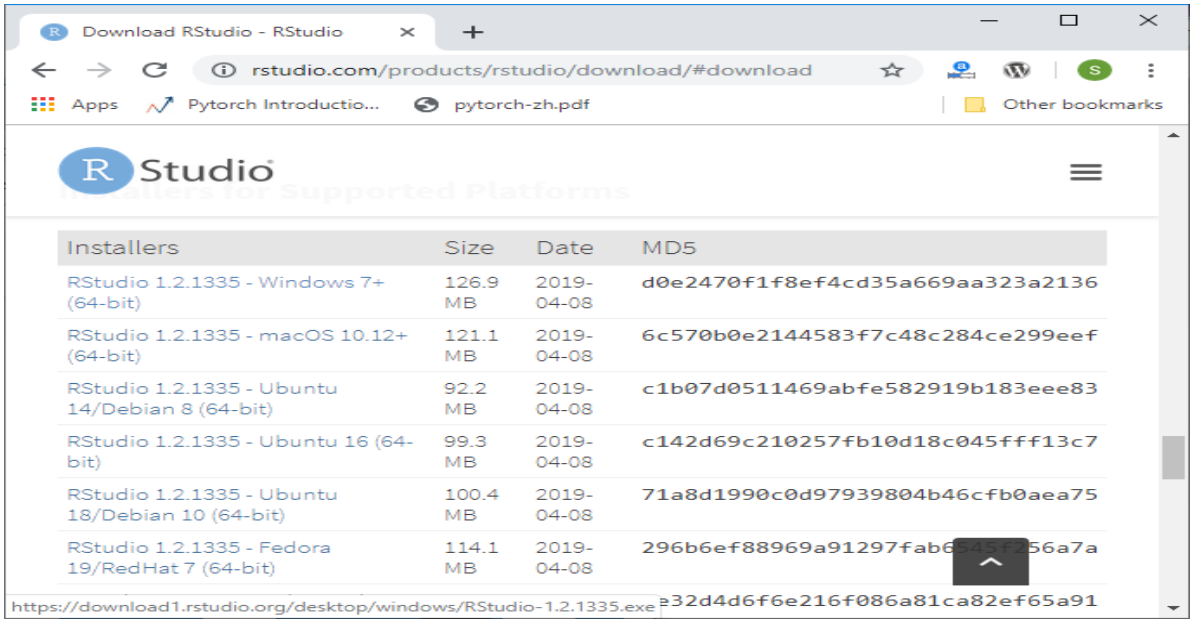
Step 2:

In the next step, we will select the RStudio desktop for open-source license and click on download.



Step 3:

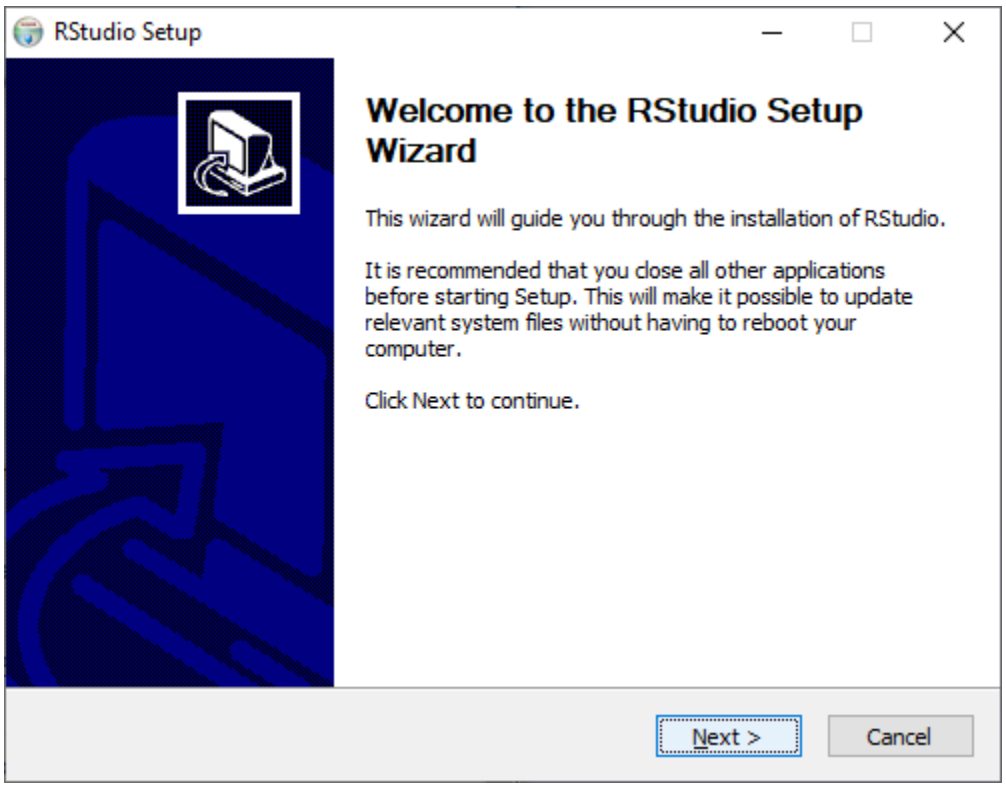
In the next step, we will select the appropriate installer. When we select the installer, our downloading of RStudion setup will start.



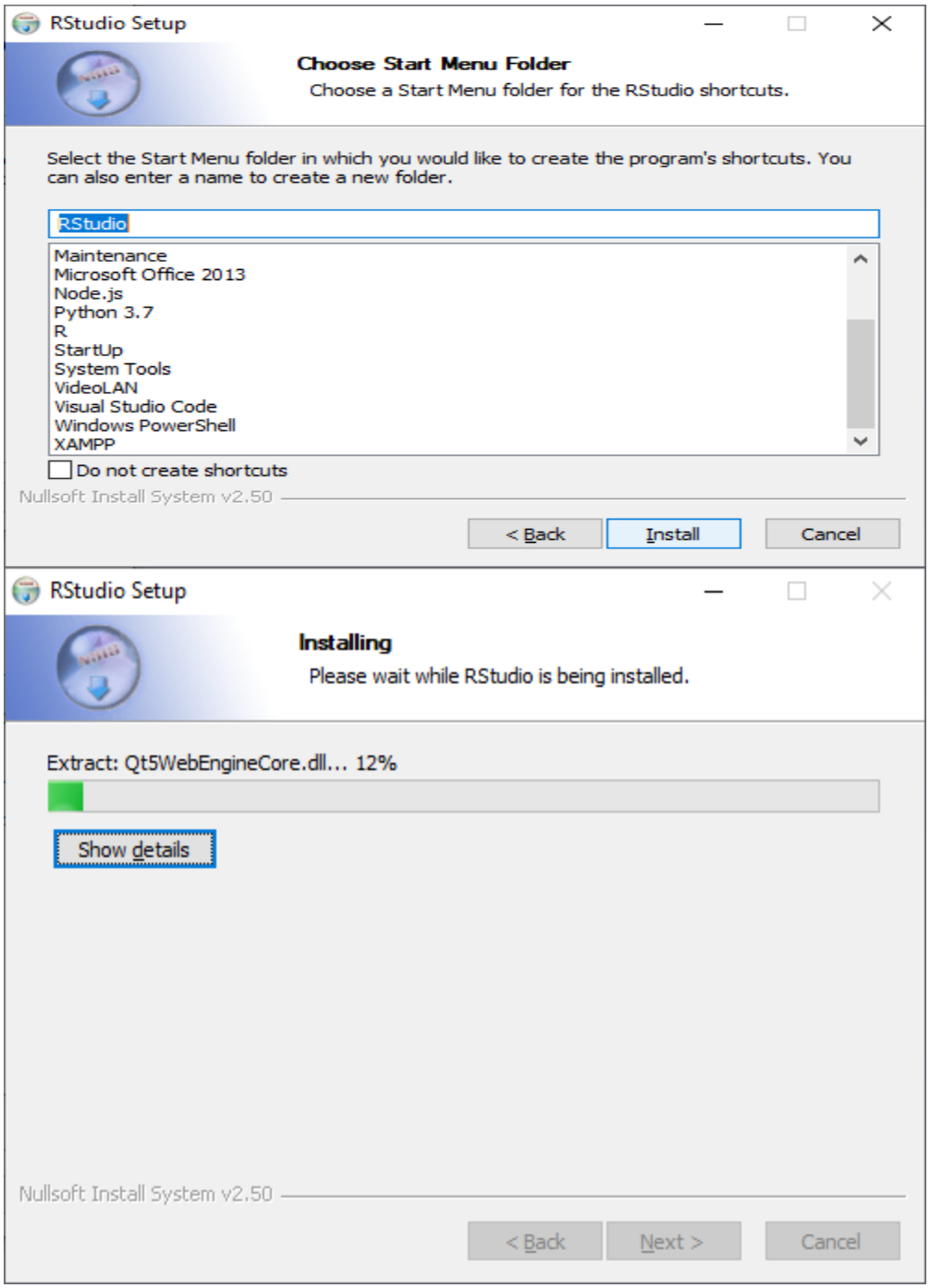
Step 4:

In the next step, we will run our setup in the following way:

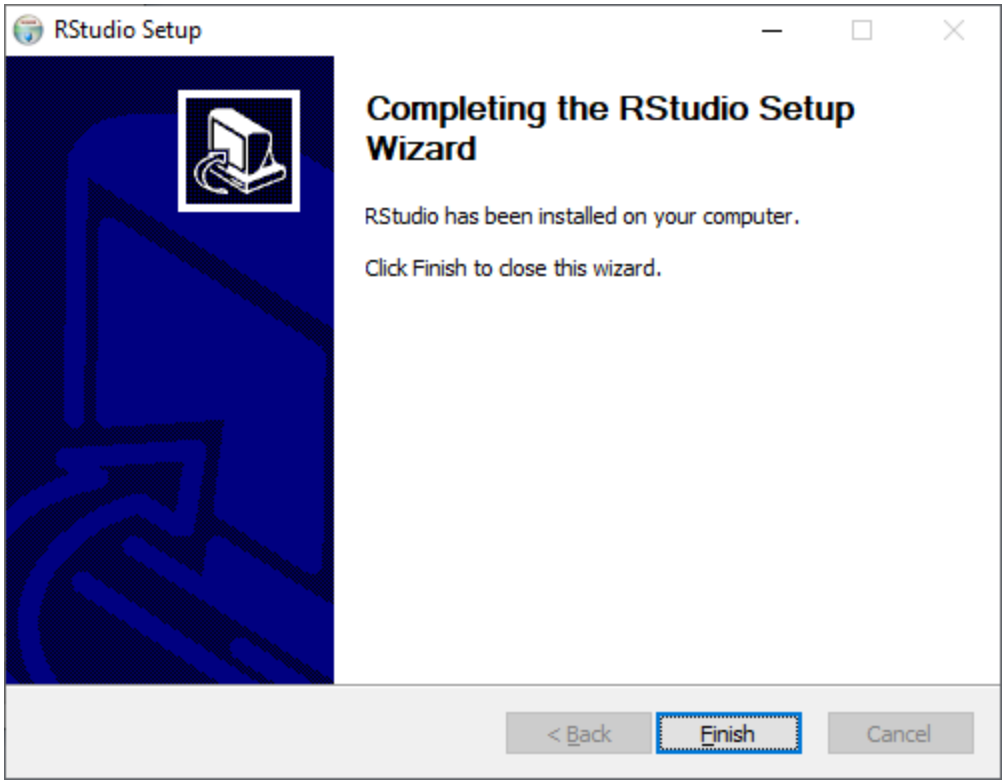
1) Click on Next.



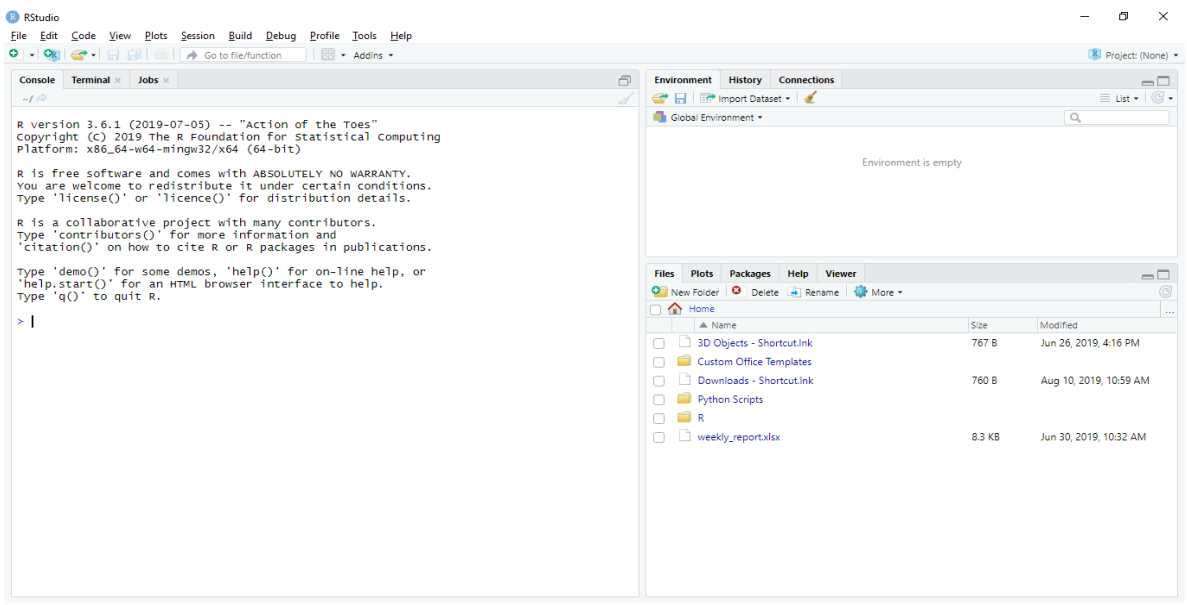
2) Click on Install.


RStudio Setup
Choose Start Menu Folder
Choose a Start Menu folder for the RStudio shortcuts.
Select the Start Menu folder in which you would like to create the program's shortcuts. You can also enter a name to create a new folder.
RStudio
Maintenance
Microsoft Office 2013
Node.js
Python 3.7
R
StartUp
System Tools
VideoLAN
Visual Studio Code
Windows PowerShell
XAMPP
Do not create shortcuts
Nullsoft Install System v2.50
< Back
Install
Cancel
RStudio Setup
Installing
Please wait while RStudio is being installed.
Extract: Qt5WebEngineCore.dll... 12%
Show details
Nullsoft Install System v2.50
< Back
Next >
Cancel

3) Click on finish.



4) RStudio is ready to work.



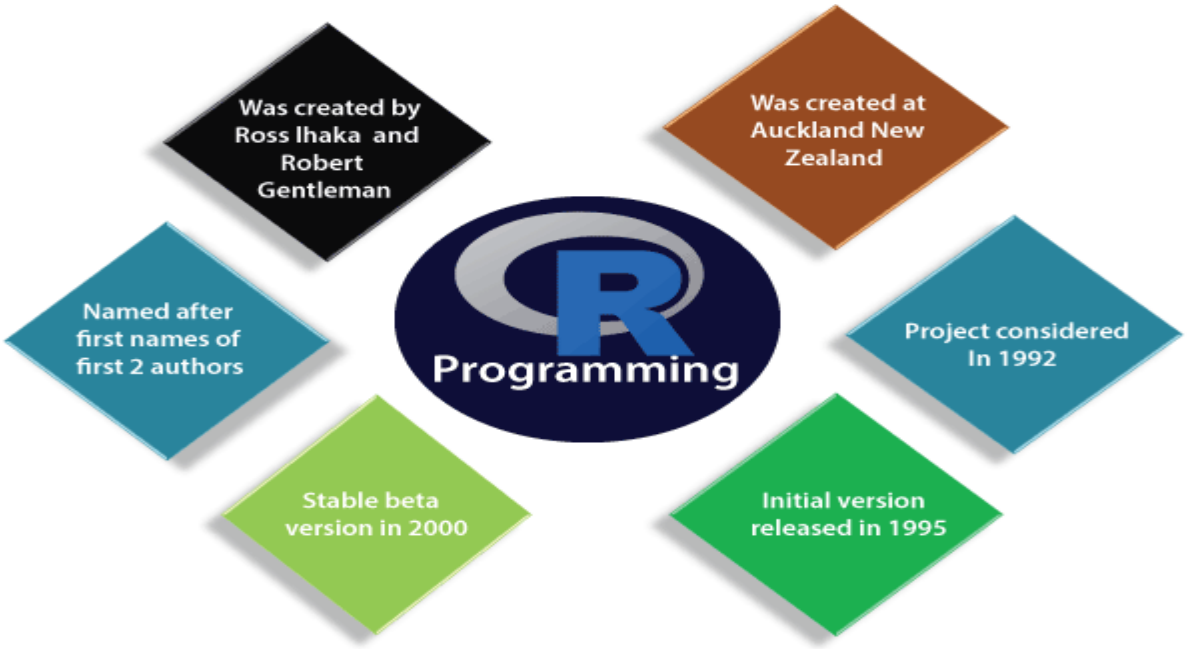
Features of R programming

R is a domain-specific programming language which aims to do data analysis. It has some unique features which make it very powerful. The most important arguably being the notation of vectors. These vectors allow us to perform a complex operation on a set of values in a single command. There are the following features of R programming:

1. It is a simple and effective programming language which has been well developed.
2. It is data analysis software.
3. It is a well-designed, easy, and effective language which has the concepts of user-defined, looping, conditional, and various I/O facilities.
4. It has a consistent and incorporated set of tools which are used for data analysis.
5. For different types of calculation on arrays, lists and vectors, R contains a suite of operators.
6. It provides effective data handling and storage facility.
7. It is an open-source, powerful, and highly extensible software.
8. It provides highly extensible graphical techniques.
9. It allows us to perform multiple calculations using vectors.
10. R is an interpreted language.

History of R Programming

11. The history of R goes back about 20-30 years ago. R was developed by Ross lhaka and Robert Gentleman in the University of Auckland, New Zealand, and the R Development Core Team currently develops it. This programming language name is taken from the name of both the developers. The first project was considered in 1992. The initial version was released in 1995, and in 2000, a stable beta version was released.



The following table shows the release date, version, and description of R language:

Version-Release	Date	Description
0.49	1997-04-23	First time R's source was released, and CRAN (Comprehensive R Archive Network) was started.
0.60	1997-12-05	R officially gets the GNU license.
0.65.1	1999-10-07	update.packages and install.packages both are included.
1.0	2000-02-29	The first production-ready version was released.
1.4	2001-12-19	First version for Mac OS is made available.
2.0	2004-10-04	The first version for Mac OS is made available.
2.1	2005-04-18	Add support for UTF-8encoding, internationalization, localization etc.

2.11	2010-04-22	Add support for Windows 64-bit systems.
2.13	2011-04-14	Added a function that rapidly converts code to byte code.
2.14	2011-10-31	Added some new packages.
2.15	2012-03-30	Improved serialization speed for long vectors.
3.0	2013-04-03	Support for larger numeric values on 64-bit systems.
3.4	2017-04-21	The just-in-time compilation (JIT) is enabled by default.
3.5	2018-04-23	Added new features such as compact internal representation of integer sequences, serialization format etc.

Why use R Programming?

- ✓ There are several tools available in the market to perform data analysis. Learning new languages is time taken. The data scientist can use two excellent tools, i.e., R and Python. We may not have time to learn them both at the time when we get started to learn data science. Learning statistical modeling and algorithm is more important than to learn a programming language. A programming language is used to compute and communicate our discovery.
- ✓ The important task in data science is the way we deal with the data: clean, feature engineering, feature selection, and import. It should be our primary focus. Data scientist job is to understand the data, manipulate it, and expose the best approach. For machine learning, the best algorithms can be implemented with R. **Keras** and **TensorFlow** allow us to create high-end machine learning techniques. R has a package to perform **Xgboost**. Xgboost is one of the best algorithms for **Kaggle competition**.

- ✓ R communicate with the other languages and possibly calls Python, Java, C++. The big data world is also accessible to R. We can connect R with different databases like **Spark** or **Hadoop**.
- ✓ In brief, R is a great tool to investigate and explore the data. The elaborate analysis such as clustering, correlation, and data reduction are done with R.

Comparison between R and Python

- ✓ Data science deals with identifying, extracting, and representing meaningful information from the data source. R, Python, SAS, SQL, Tableau, MATLAB, etc. are the most useful tools for data science. R and Python are the most used ones. But still, it becomes confusing to choose the better or the most suitable one among the two, R and Python.

Comparison Index	R	Python
Overview	"R is an interpreted computer programming language which was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand ." The R Development Core Team currently develops R. R is also a software environment which is used to analyze statistical information, graphical representation, reporting, and data modeling.	Python is an Interpreted high-level programming language used for general-purpose programming. Guido Van Rossum created it, and it was first released in 1991. Python has a very simple and clean code syntax. It emphasizes the code readability and debugging is also simple and easier in Python.
Specialties for data science	R packages have advanced techniques which are very useful for statistical work. The CRAN text view is provided by many useful R packages. These packages cover everything from Psychometrics to Genetics to Finance.	For finding outliers in a data set both R and Python are equally good. But for developing a web service to allow peoples to upload datasets and find outliers, Python is better.

Functionalities	For data analysis, R has inbuilt functionalities	Most of the data analysis functionalities are not inbuilt. They are available through packages like Numpy and Pandas
Key domains of application	Data visualization is a key aspect of analysis. R packages such as ggplot2, ggvis, lattice, etc. make data visualization easier.	Python is better for deep learning because Python packages such as Caffe, Keras, OpenNN, etc. allows the development of the deep neural network in a very simple way.
Availability of packages	There are hundreds of packages and ways to accomplish needful data science tasks.	Python has few main packages such as viz, Sccikit learn, and Pandas for data analysis of machine learning, respectively.

Applications of R

There are several-applications available in real-time. Some of the popular applications are as follows:

- Facebook
- Google
- Twitter
- HRDAG
- Sunlight Foundation
- RealClimate
- NDAA
- XBOX ONE
- ANZ
- FDA

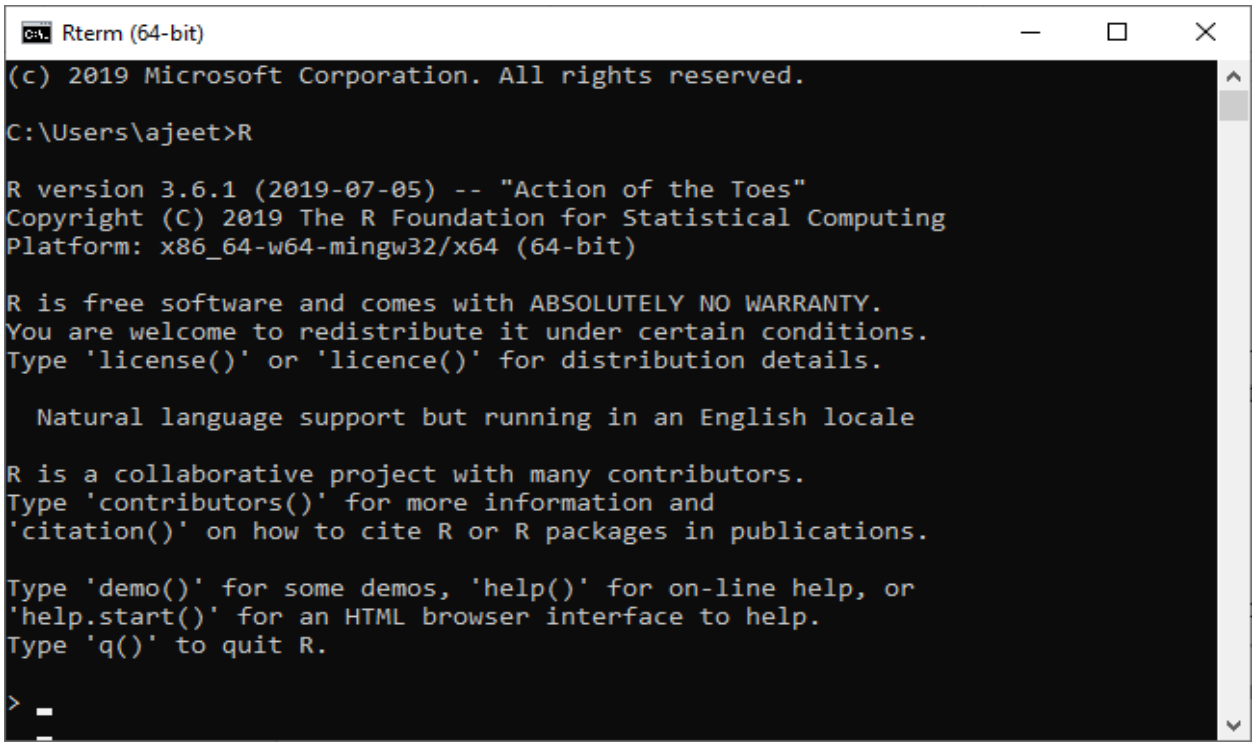
Syntax of R Programming

- ✓ R Programming is a very popular programming language which is broadly used in data analysis. The way in which we define its code is quite simple. The "Hello

World!" is the basic program for all the languages, and now we will understand the syntax of R programming with "Hello world" program. We can write our code either in command prompt, or we can use an R script file.

R Command Prompt

- ✓ It is required that we have already installed the R environment set up in our system to work on the R command prompt. After the installation of R environment setup, we can easily start R command prompt by typing R in our Windows command prompt. When we press enter after typing R, it will launch interpreter, and we will get a prompt on which we can code our program.

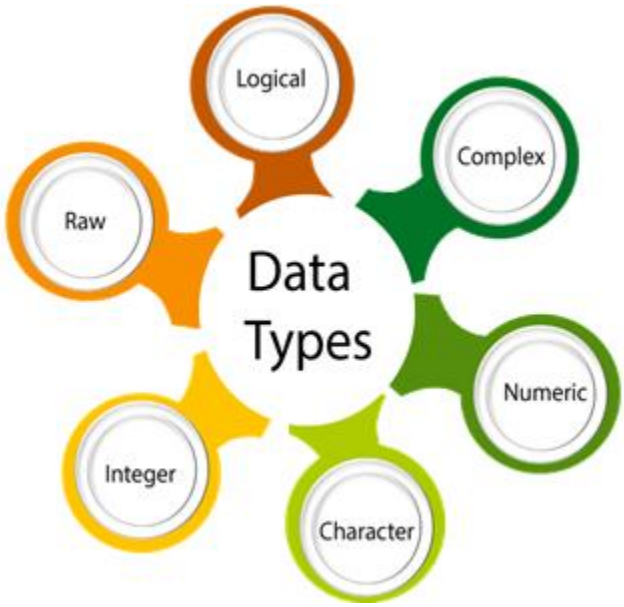


Hello, World!" Program

The code of "Hello World!" in R programming can be written as:

Data Types in R Programming

- ✓ In programming languages, we need to use various variables to store various information. Variables are the reserved memory location to store values. As we create a variable in our program, some space is reserved in memory.
- ✓ In R, there are several data types such as integer, string, etc. The operating system allocates memory based on the data type of the variable and decides what can be stored in the reserved memory.
- ✓ There are the following data types which are used in R programming:



Data type	Example	Description
Logical	True, False	It is a special data type for data with only two possible values which can be construed as true/false.
Numeric	12,32,112,5432	Decimal value is called numeric in R, and it is the default computational data type.
Integer	3L, 66L, 2346L	Here, L tells R to store the value as an integer,
Complex	Z=1+2i, t=7+3i	A complex value in R is defined as the pure imaginary value i.
Character	'a', '"good"', "TRUE", '35.4'	In R programming, a character is used to represent string values. We convert objects into character values with the help ofas.character() function.
Raw		A raw data type is used to holds raw bytes.

Data type :

- ✓ A variable can store different types of values such as numbers, characters etc. These different types of data that we can use in our code are called **data types**. For example,

```
x <- 123L
```

Here, `123L` is an integer data. So the data type of the variable `x` is `integer`.

We can verify this by printing the class of `x`.

```
x <- 123L



# print value of x
X<-123
print(x)

# print type of x
print(class(x))
```

Output

```
[1] 123
[1] "integer"
```

Here, `x` is a variable of data type `integer`.

Different Types of Data Types

In R, there are 6 basic data types:

- `logical`
- `numeric`
- `integer`
- `complex`
- `character`
- `raw`

1. Logical Data Type

The `logical` data type in R is also known as **boolean** data type. It can only have two values: `TRUE` and `FALSE`. For example,

```
bool1 <- TRUE

print(bool1)
print(class(bool1))

bool2 <- FALSE

print(bool2)
print(class(bool2))
```

Output

```
[1] TRUE
[1] "logical"
[1] FALSE
[1] "logical"
```

In the above example,

- *bool1* has the value `TRUE`,
- *bool2* has the value `FALSE`.

Here, we get `"logical"` when we check the type of both variables.

Note: You can also define logical variables with a single letter - `T` for `TRUE` or `F` for `FALSE`. For example,

```
is_weekend <- F
print(class(is_weekend))  # "logical"
```

2. Numeric Data Type

- In R, the `numeric` data type represents all real numbers with or without decimal values. For example,

```
# floating point values
weight <- 63.5

print(weight)
print(class(weight))

# real numbers
height <- 182

print(height)
print(class(height))
```

Output

```
[1] 63.5
[1] "numeric"
[1] 182
[1] "numeric"
```

- Here, both *weight* and *height* are variables of `numeric` type.

3. Integer Data Type

- The `integer` data type specifies real values without decimal points. We use the suffix `L` to specify integer data. For example,

```
integer_variable <- 186L
print(class(integer_variable))
```

Output

```
[1] "integer"
```

- Here, `186L` is an integer data. So we get `"integer"` when we print the class of *integer_variable*.

4. Complex Data Type

- The `complex` data type is used to specify purely imaginary values in R. We use the suffix `i` to specify the imaginary part. For example,

```
# 2i represents imaginary part
complex_value <- 3 + 2i



# print class of complex_value
print(class(complex_value))
```

Output

```
[1] "complex"
```

Here, `3 + 2i` is of `complex` data type because it has an imaginary part `2i`.

5. Character Data Type

- The `character` data type is used to specify character or string values in a variable.
- In programming, a string is a set of characters. For example, `'A'` is a single character and `"Apple"` is a string.
- You can use single quotes `''` or double quotes `""` to represent strings. In general, we use:

- `''` for character variables
- `""` for string variables

For example,

```
# create a string variable
fruit <- "Apple"

print(class(fruit))

# create a character variable
my_char <- 'A'

print(class(my_char))
```

Output

```
[1] "character"
[1] "character"
```

Here, both the variables - *fruit* and *my_char* - are of `character` data type.

6. Raw Data Type

A `raw` data type specifies values as raw bytes. You can use the following methods to convert character data types to a raw data type and vice-versa:

- `charToRaw()` - converts character data to raw data
- `rawToChar()` - converts raw data to character data

For example,

```
# convert character to raw
raw_variable <- charToRaw("Welcome ")


print(raw_variable)
print(class(raw_variable))
```

```
# convert raw to character
char_variable <- rawToChar(raw_variable)



print(char_variable)
print(class(char_variable))
```

```
# convert character to raw

raw_variable <- charToRaw("Welcome")

print(raw_variable)

print(class(raw_variable))

raw_variable <- rawToChar(raw_variable)

print(raw_variable)
```

Output

```
[1] 57 65 6c 63 6f 6d 65 20 74 6f 20 50 72 6f 67 72 61 6d 69 7a
[1] "raw"
[1] "Welcome "
[1] "character"
```

In this program,

- We have first used the `charToRaw()` function to convert the string `"Welcome to Programming"` to raw bytes.

  This is why we get `"raw"` as output when we print the class of *raw_variable*.
- Then, we have used the `rawToChar()` function to convert the data in *raw_variable* back to character form.

This is why we get "character" as output when we print the class of *char_variable*.

Basic Programs

How to the user input in 'R'

There are two methods in R.

- Using **readline()** method
- Using **scan() method**

Using readline() method

- In R language **readline()** method takes input in string format. If one inputs an integer then it is inputted as a string, lets say, one wants to input **255**, then it will input as **"255"**, like a string.
- So one needs to convert that inputted value to the format that he needs. In this case, string **"255"** is converted to integer 255. To convert the inputted value to the desired data type, there are some functions in R,

- **as.integer(n)**; —> convert to integer
- **as.numeric(n)**; —> convert to numeric type (float, double etc)
- **as.complex(n)**; —> convert to complex number (i.e 3+2i)
- **as.Date(n)** —> convert to date …, etc

Syntax:

var = readline();

var = as.integer(var);

Note that one can use "<-" instead of "="