

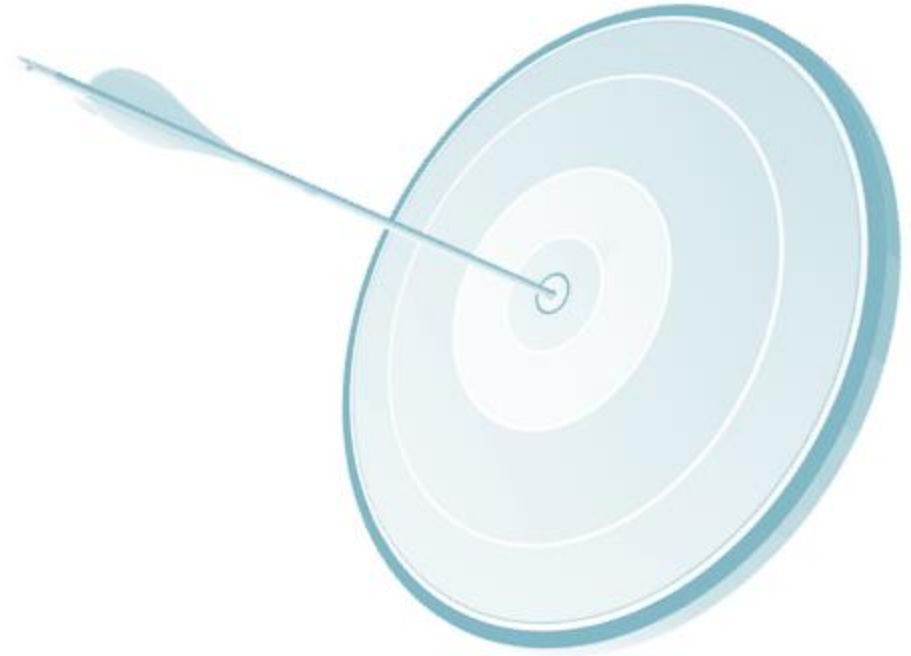
LINEAR REGRESSION WITH R

View Business Analytics with R course details at www.edureka.co/r-for-analytics

Objectives

At the end of this session, you will be able to

- What is data mining
- What is Business Analytics
- Stages of Analytics / data mining
- What is R
- Overview of Machine Learning
- What is Linear Regression
- Case Study



- Generally, data mining is the process of studying data from maximum possible dimensions and summarizing it into useful information
- Technically, data mining is the process of finding correlations or patterns among dozens of fields in large data generated from business
- Or you can say, data mining is the process finding useful information from the data and then devising knowledge out of it for improving future of our business

» Data ??

Data are any facts, numbers, or text is getting produced by existing system


» Information ??

The patterns, associations, or relationships among all this data can provide information


» Knowledge ??

Information can be converted into knowledge about historical patterns and future trends. For example summary of sales in off season may help to start some offers in that period to increase sales

Why Business Analytics is getting popular these days ?



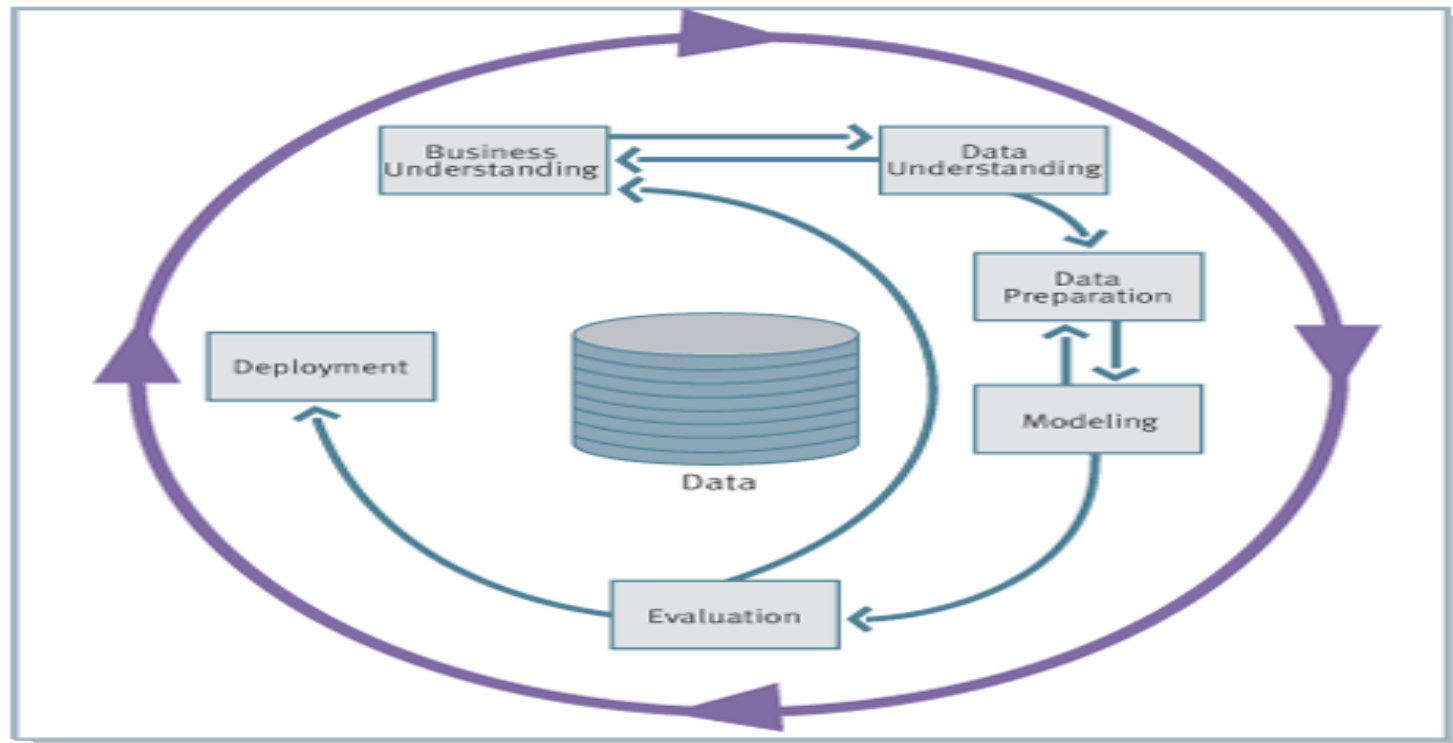
Cost of storing data



Cost of processing data

Stages of Analytics / Data Mining

Cross Industry standard Process for data mining (CRISP – DM)



What is R

R is Programming Language

R is Environment for Statistical Analysis

R is Data Analysis Software

```
R Console

Error in library(ggplot2) : there is no package called 'ggplot2'
> install.packages("ggplot2")
Installing package into 'C:/Users/User/Documents/R/win-library/3.0'
(as 'lib' is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.0/ggplot2_1.0.0.zip'
Content type 'application/zip' length 2672889 bytes (2.5 Mb)
opened URL
downloaded 2.5 Mb

package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\User\AppData\Local\Temp\Rtmp4M0fE3\downloaded_packages
> library(ggplot2)
Warning message:
package 'ggplot2' was built under R version 3.0.3
> x<- runif(50)
> y<- runif(50)
> D<- cbind(x,y)
> plot(D)
> km<- kmeans(D,4)
> str(km)
List of 7
 $ cluster      : int [1:50] 1 2 2 3 2 1 4 1 1 1 ...
 $ centers      : num [1:4, 1:2] 0.336 0.68 0.128 0.921 0.242 ...
 .. attr(*, "dimnames")=List of 2
 .. ..$ : chr [1:4] "1" "2" "3" "4"
 .. ..$ : chr [1:2] "x" "y"
 $ totss       : num 8.87
 $ withinss    : num [1:4] 0.522 0.592 0.239 0.378
 $ tot.withinss: num 1.73
 $ betweenss   : num 7.14
 $ size        : int [1:4] 14 17 8 11
 - attr(*, "class")= chr "kmeans"
> |
```

- Effective and fast data handling and storage facility
- A bunch of operators for calculations on arrays, lists, vectors etc
- A large integrated collection of tools for data analysis, and visualization
- Facilities for data analysis using graphs and display either directly at the computer or paper
- A well implemented and effective programming language called 'S' on top of which R is built
- A complete range of packages to extend and enrich the functionality of R

Data Visualization in R

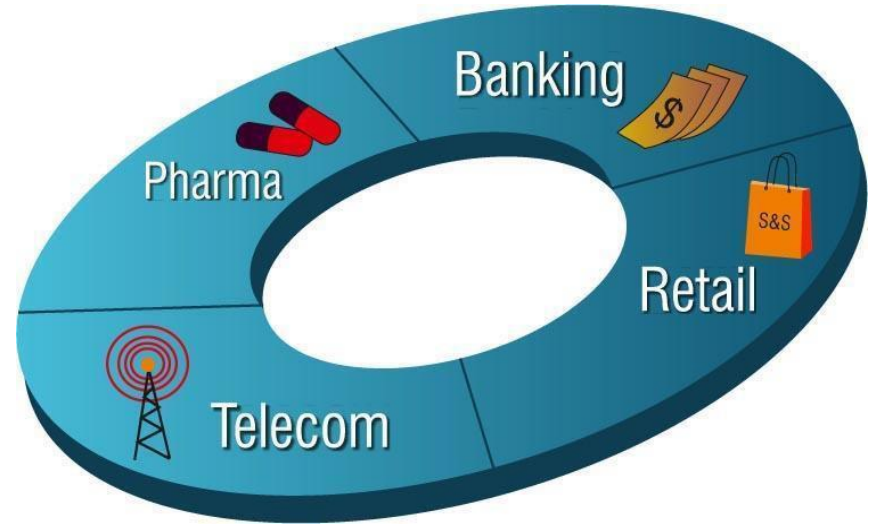
→ This plot represents the locations of all the traffic signals in the city.

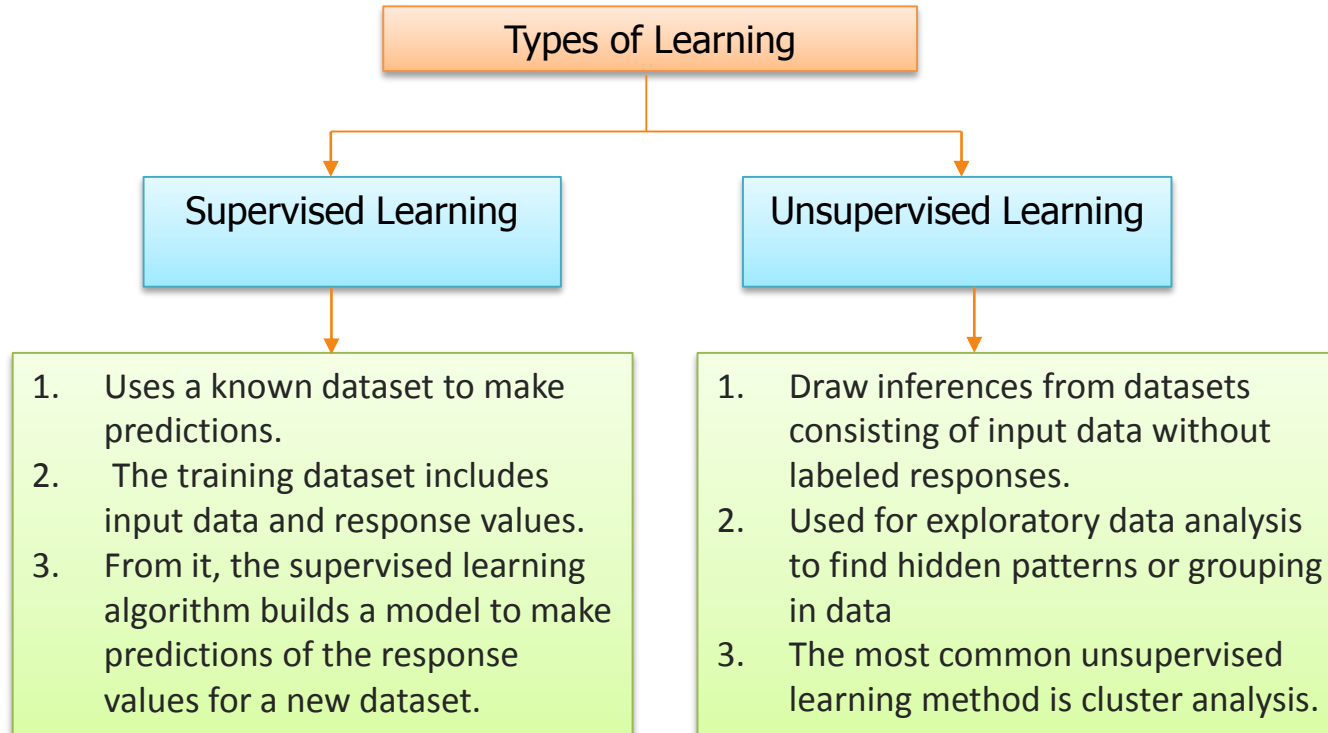
→ It is recognizable as Toronto without any other geographic data being plotted - the structure of the city comes out in the data alone.

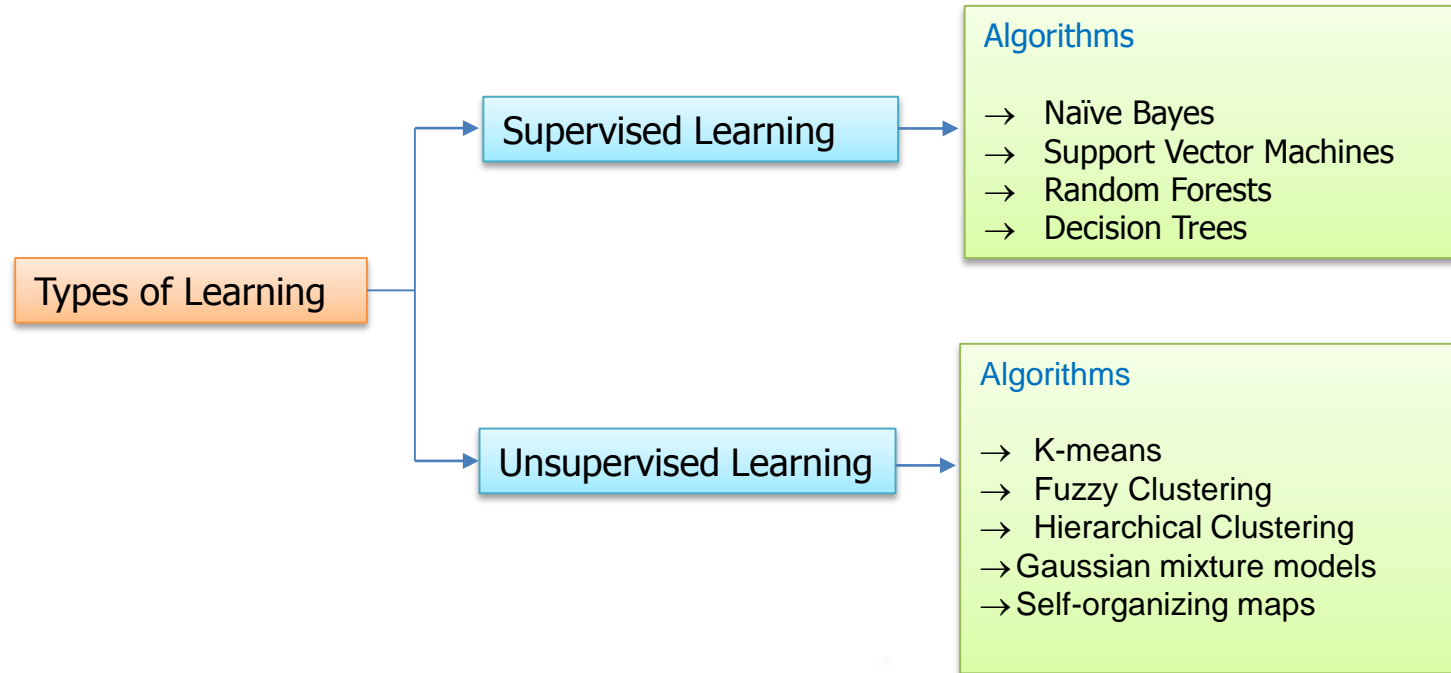


Who Uses R : Domains

- Telecom
- Pharmaceuticals
- Financial Services
- Life Sciences
- Education, etc







Linear Regression

What is Linear Regression??

- In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variable) denoted X .
- The case of one explanatory variable is called simple linear regression.
- For more than one explanatory variable, the process is called multiple linear regression
- Data are modeled using linear predictor functions, and unknown model parameters are estimated from the data.

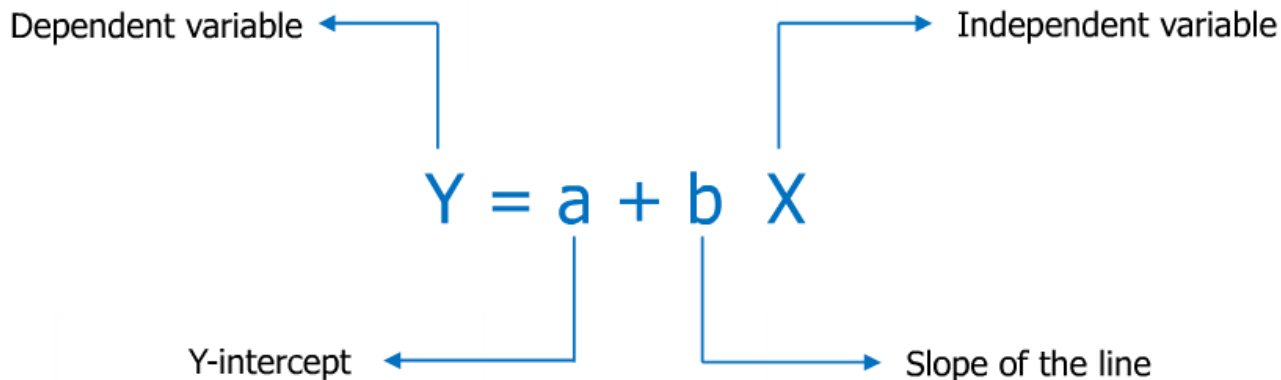
Where to Use Linear Regression??

Linear regression has many practical uses.

Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or reduction, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y , the fitted model can be used to make a prediction of the value of y .
- Given a variable y and a number of variables X_1, \dots, X_p that may be related to y , linear regression analysis can be applied to quantify the strength of the relationship between y and the X_j , to assess which X_j may have no relationship with y at all, and to identify which subsets of the X_j contain redundant information about y .

Equation - Linear Regression??



Y-intercept (a) is that value of the Dependent Variable(y) when the value of the Independent Variable(x) is zero. It is the point at which the line cuts the y-axis.

Slope (b) is the change in the Dependent Variable for a unit increase in the Independent Variable. It is the tangent of the angle made by the line with the x-axis.

Linear Regression Case-study

Problem Statement

Computer manufacturing company is trying to analyse the data of the price of a computer with another independent variable like- Cpu speed, Hard disc, RAM, Screen Size, CD (yes/no), produced by premium company(yes/no) and so on. Based on this data, company wants to decide on the price of a new configuration of PC.

The company dataset looks like this:

This problem can be solved by a linear regression model

The Computer_Data looks like this:

	price	speed	hd	ram	screen	cd	multi	premium	ads	trend
1	1499	25	80	4	14	no	no	yes	94	1
2	1795	33	85	2	14	no	no	yes	94	1
3	1595	25	170	4	15	no	no	yes	94	1
4	1849	25	170	8	14	no	no	no	94	1
5	3295	33	340	16	14	no	no	yes	94	1
6	3695	66	340	16	14	no	no	yes	94	1
7	1720	25	170	4	14	yes	no	yes	94	1
8	1995	50	85	2	14	no	no	yes	94	1
9	2225	50	210	8	14	no	no	yes	94	1
10	2575	50	210	4	15	no	no	yes	94	1

Demo

More Information on R setup and applications at:

<http://www.edureka.in/blog/category/business-analytics-with-r/>

Course Topics

- **Module 1**
 - » Introduction to Business Analytics
- **Module 2**
 - » Introduction to R Programming
- **Module 3**
 - » Data Manipulation in R
- **Module 4**
 - » Data Import Techniques in R
- **Module 5**
 - » Exploratory Data Analysis
- **Module 6**
 - » Data Visualization in R
- **Module 7**
 - » Data mining: Clustering Techniques
- **Module 8**
 - » Data Mining: Association rule mining and Sentiment analysis
- **Module 9**
 - » Linear and Logistic Regression
- **Module 10**
 - » Anova and Predictive Analysis
- **Module 11**
 - » Data Mining: Decision Trees and Random forest
- **Module 12**
 - » Final Project Business Analytics with R class – Census Data

Thank you!

