DATA ANALYSIS USING PYTHON

FACULTY DEVLOPMENT PROGRAM ON PYTHON TRAINING

Organized BY

Department of Statistics

Osmania University

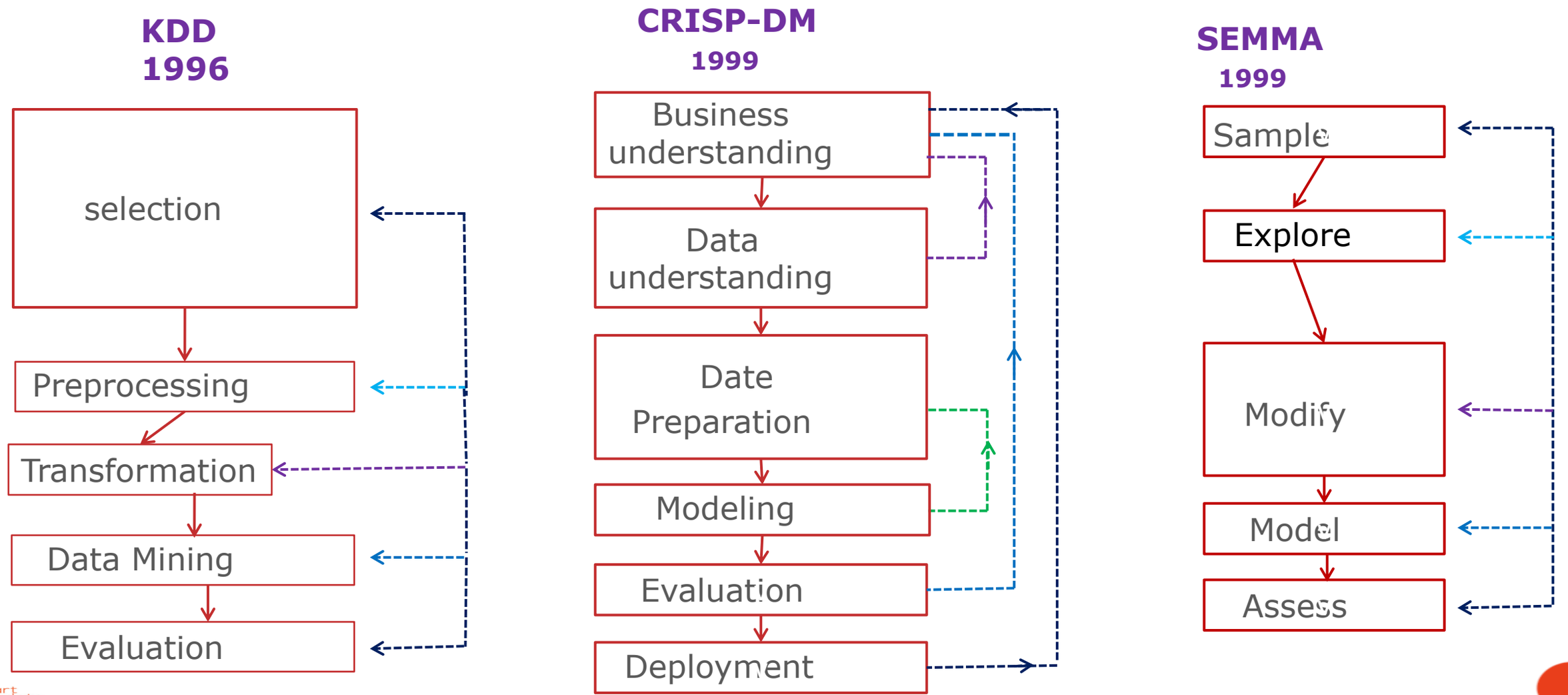
Contact Person: Dr. Manneni Venu Gopala Rao
Ph. No: 9866633975
Company: Juxt-Smartmandate Analytical Solutions Pvt. Ltd.
Email: venugopal@jsm.email

Agenda

- Overview of Data Mining/ Analysis Process
- Overview of Python Language
  - The Present state
  - Benefits over other Tools
  - Some of the applications
- Understanding standard Libraries in Python
  - Review of some of the important libraries required for Data Analysis
    - Numpy - Python for numerical computations
    - Pandas - Understanding data using Descriptive statistics
    - Matplotlib – Understanding data with Visz
    - Scipy – Inferential Statistics
    - Sckit Learn – Data Preparation and Modelling
- Case study1 – An application of Regression
- Case study2 - An application of Classification

Overview of Data Mining Process

Summary of data Mining Frameworks



Steps in Analytical Process

Business Understanding

Data Understanding

Data preparation

Modelling

Evaluation

Deployment

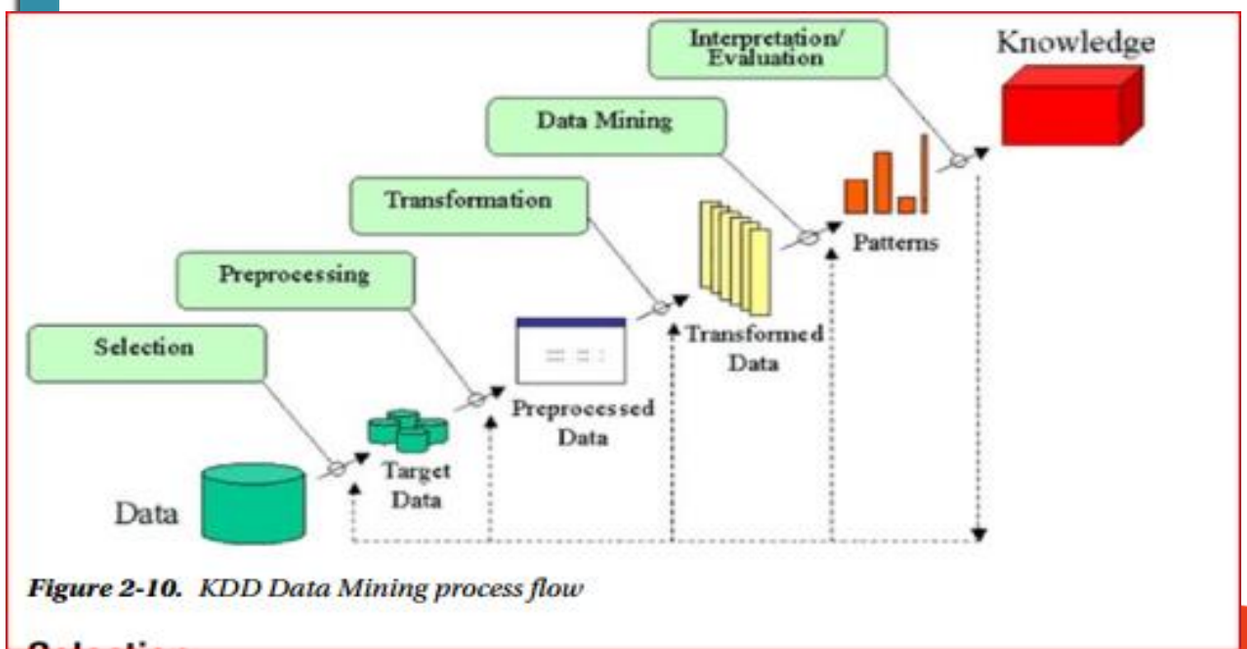


Figure 2-10. KDD Data Mining process flow

Steps in Analytical Process

1. Business Understanding

- Identify the Business objective
- Assess the situation
- Determine the Analytical goals
- Produce a project plan

2. Data Understanding

- ✓ Collect the data
- ✓ Describe the data
- ✓ Explore the data
- ✓ Verify the data Quality

3. Data Preparation

- Select the data
- Clean the data
- Construct the data
- Integrate the data
- Format the data

4. Modeling

- Select a modeling technique
- Generate a test Design
- Build a model
- Assess a model

5. Evaluation

- ✓ Evaluate the results
- ✓ Review the process
- ✓ Determine the next steps

6. Deployment

- Deploying the plan
- Monitoring and maintenance of the plan
- Producing the final report
- Reviewing the project

Typical Effort for each Process

- Business Understanding >> 5 to 15 %
- Data Understanding >> 5 to 10 %
- Data Preparation >> 50 to 60 %
- Modeling >> 5 to 15 %
- Evaluation >> 5 to 10 %
- Deployment >> 10 to 15 %

In Action, we follow

- Business Understanding: More of a Domain Expert problem
- Data Understanding
  - Data Exploration (check for any outliers / missing values /…)
    - Understanding using Descriptive Statistics / Correlations… Etc.
    - Understanding data with Visualizations – Histograms / Box Plots / Scatter plots / Correlation plots…

Data understanding…

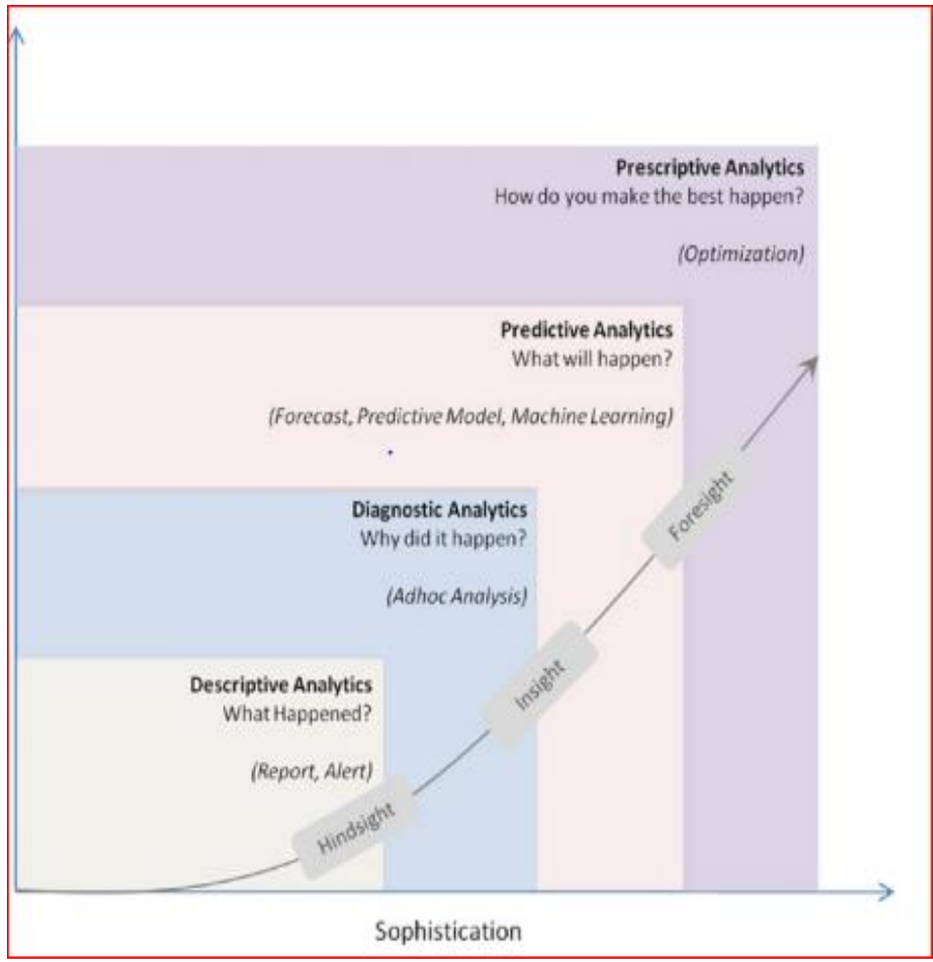
- Data Pre-processing
  - Use different data transformations in order to expose the structure of prediction problems in a better way
    - Standardizing
    - Normalizing etc.

Modelling

- Based on the type of data, we need to select the modelling algorithm, whether it is a
  - Regression Problem
  - Cluster Analysis
  - Classification Analysis
- Statistical Models are mainly divided into 3 different types like,
  - Grouping / Predicting / Association

    (based on the type of dependent/Independent variables, different methods exist and, we need to choose accordingly)

Analytics Continuum



Source: Gartner

Overview of Python

The programmers today use Python as it has created a mark for itself in the software development with characteristic features like-

- Interactive
- Interpreted
- Modular
- Dynamic
- Object-oriented
- Portable
- High level
- Extensible in C++ & C

Advantages or Benefits of Python

Extensive Support Libraries - Numpy/Pandas/ Sckit learn /NLP/Keras/Tensor Flow

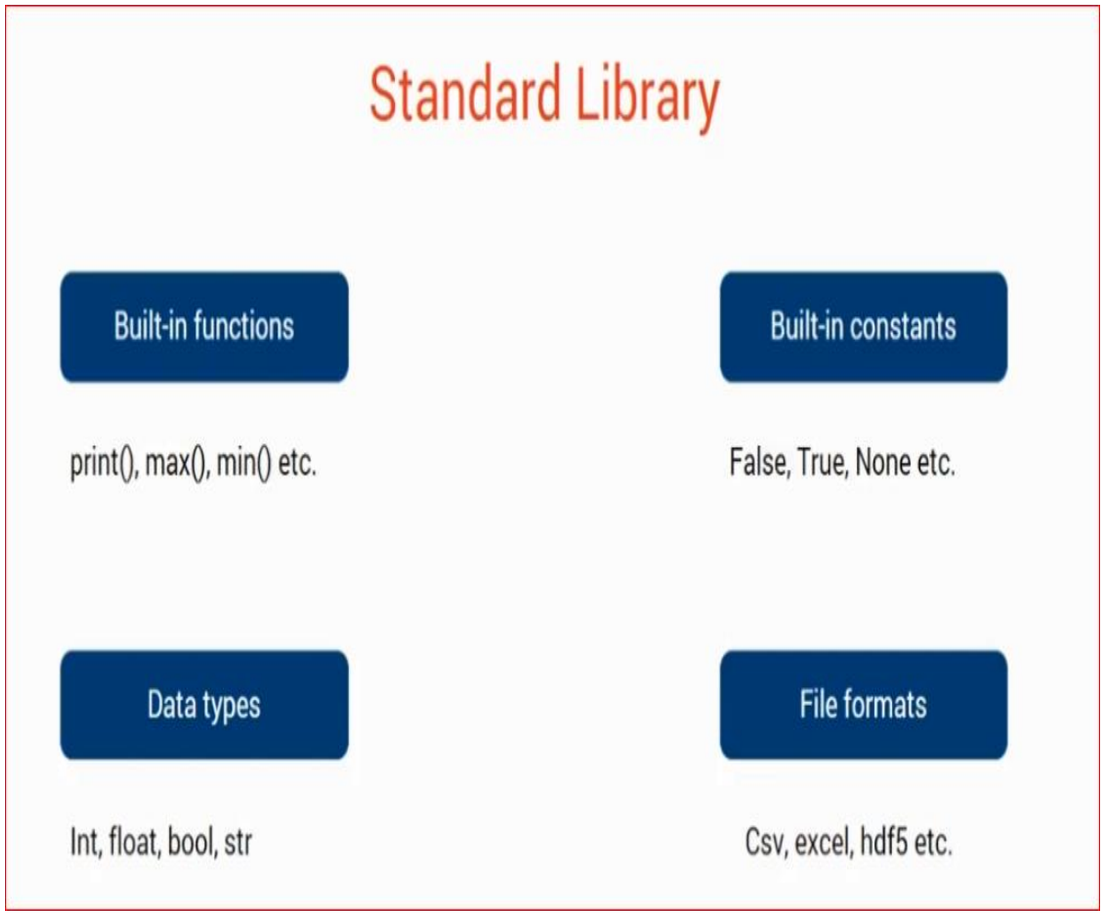
Open Source/ Huge Community
Lot of community help in terms of code and development

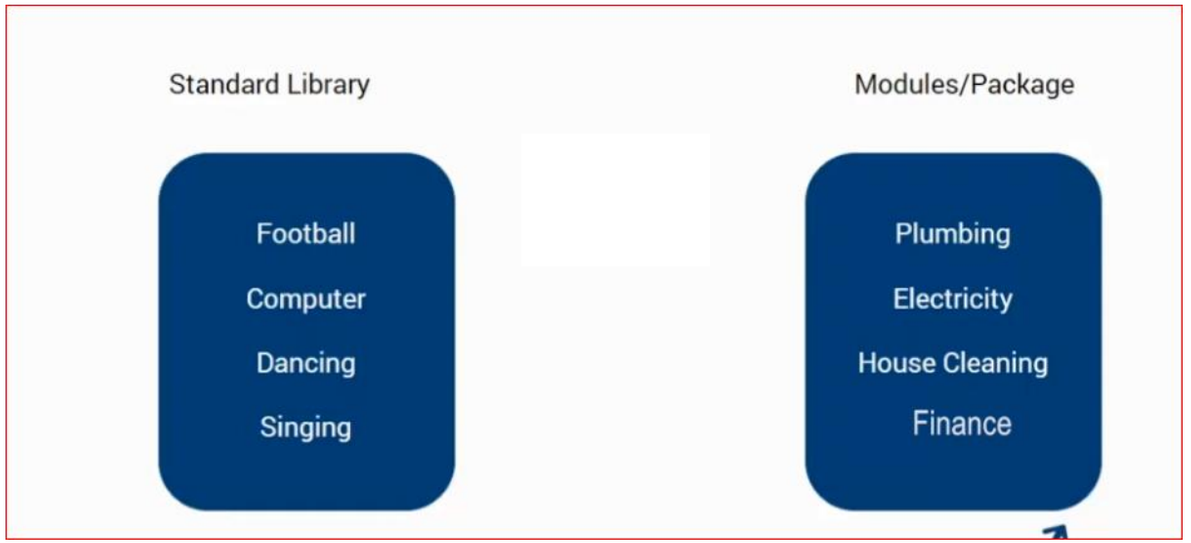
Integration Feature/Productivity

Python integrates the Enterprise Application Integration that makes it easy to develop Web services by invoking COM or COBRA components.

Python powers Django, a complete and open source web application framework. Frameworks - like Ruby on Rails - can be used to simplify the development process.

Understanding Standard Libraries in Python



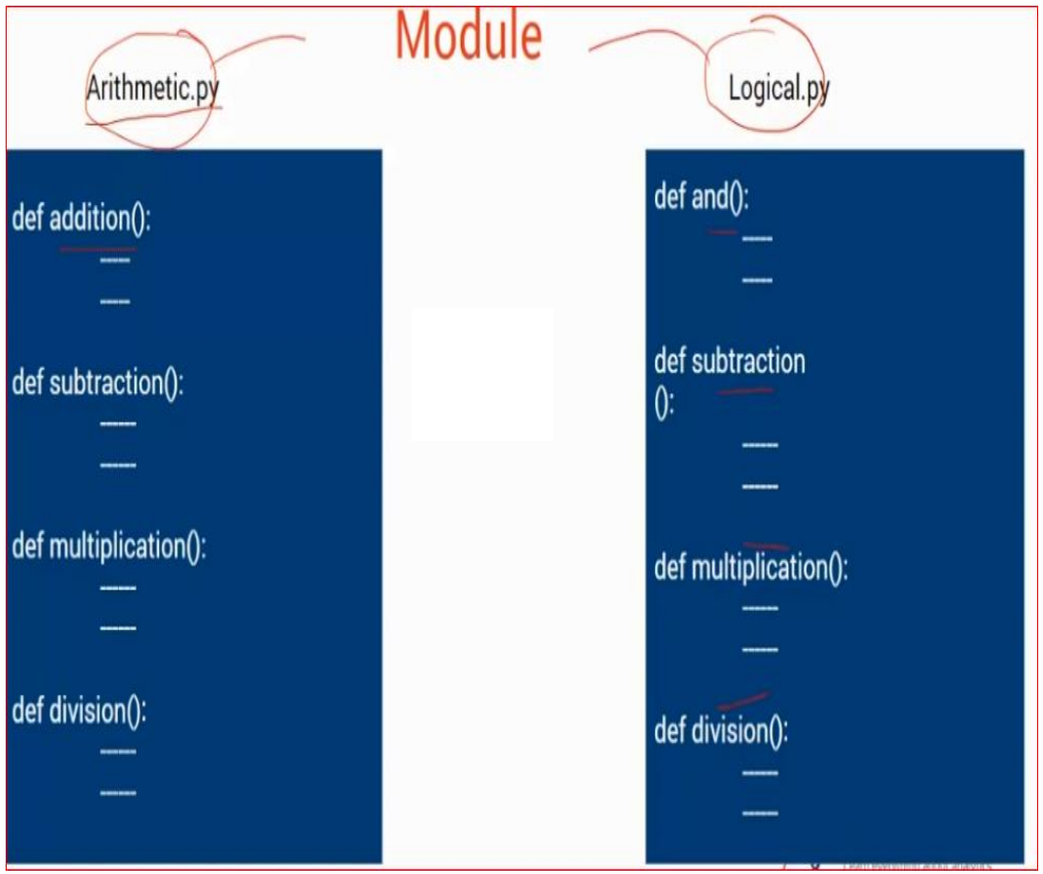


Python has some standard library which python itself uses internally but it depends on Modules/ Packages for some of the tasks, which it is not good at.

As being open source, lot of people designed modules/ packages in python for various tasks, which people will use for their work

Module/Package (Library)

Module is a collection of functions



Package is a collection of modules

Once we install the packages, we will import modules from the packages/ Libraries

Some of the more useful Libraries/ Packages for Data Analysis

- Numpy – Numerical Python
- Pandas - Python for data Analysis
- Matplotlib – Viszvalization
- SciPy - Inferential statistics
- Sckit- learn/ Sklearn - Data Preparation and Machine Learning Modelling
- NLTK/Gensim/Spacy – Natural Language Processing (Text Analytics)
- Keras/Tensorflow/… - Deep Learning Applications (NN/ Computer vision/ Chat bots…etc)

Numpy – Arrays/ Matrices with same data types

- NumPy stands for Numerical Python.
- This library contains basic linear algebra functions, fourier transformations, advanced random number capabilities.
- NumPy works with Python objects called multi-dimensional **arrays**.
- Arrays are basically collections of values, and they have one or more dimensions.
- NumPy array data structure is also called *ndarray* (n-dimensional array).
- An array with one dimension is called a **vector** and an array with two dimensions is called a **matrix**.

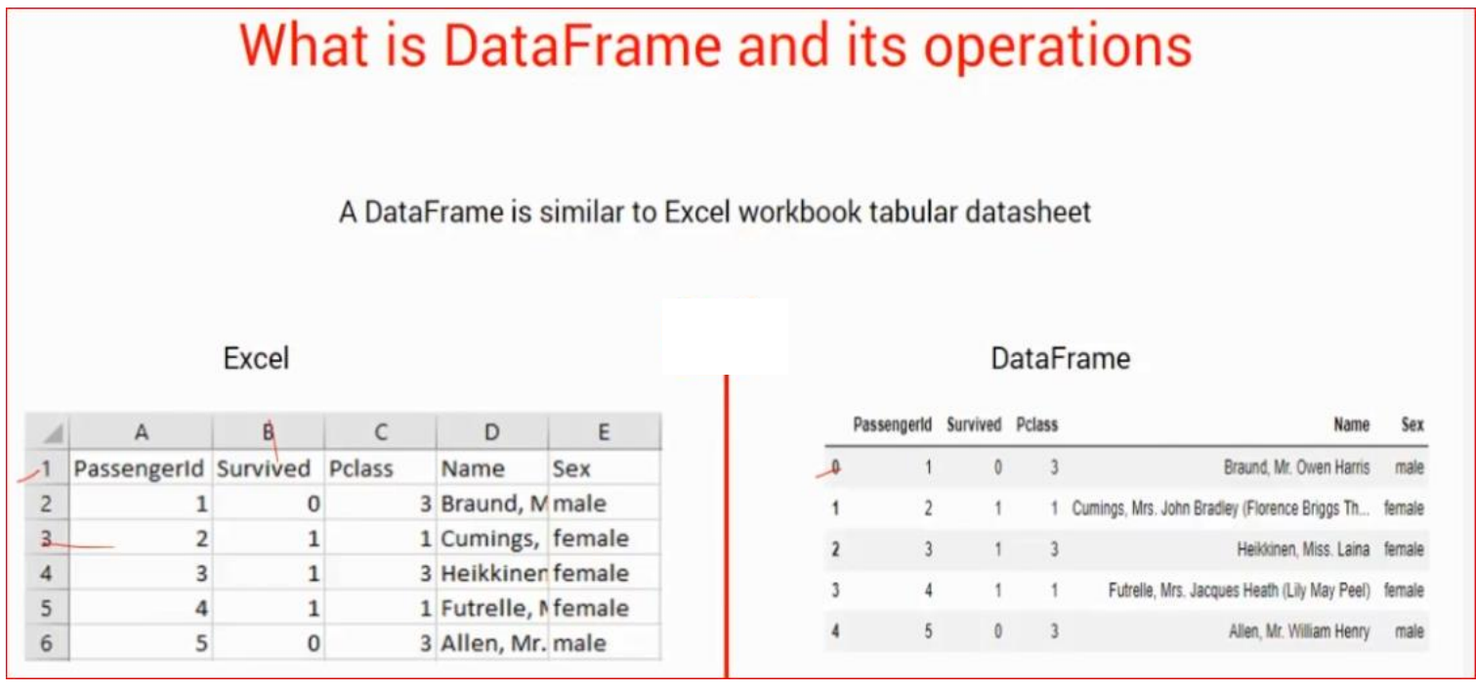
NumPy Contd…

- With NumPy we can perform the below
  - Working and Inspecting Arrays
  - Indexing and Slicing
  - Sorting and Reshaping
  - Combining and Splitting
  - Adding and Removing elements
  - Descriptive Statistics
  - Doing Math
  - Exporting numpy data into delimited files

..\Codes\numpy

Pandas

Pandas stands for “Python for Data Analysis” library. The name is derived from the term “panel data”, an econometrics term for multidimensional structured data sets.



	A	B	C	D	E
1	PassengerId	Survived	Pclass	Name	Sex
2	1	0	3	Braund, M	male
3	2	1	1	Cumings,	female
4	3	1	3	Heikkinen	female
5	4	1	1	Futrelle, M	female
6	5	0	3	Allen, Mr.	male

	PassengerId	Survived	Pclass	Name	Sex
0	1	0	3	Braund, Mr. Owen Harris	male
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female
2	3	1	3	Heikkinen, Miss. Laina	female
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
4	5	0	3	Allen, Mr. William Henry	male

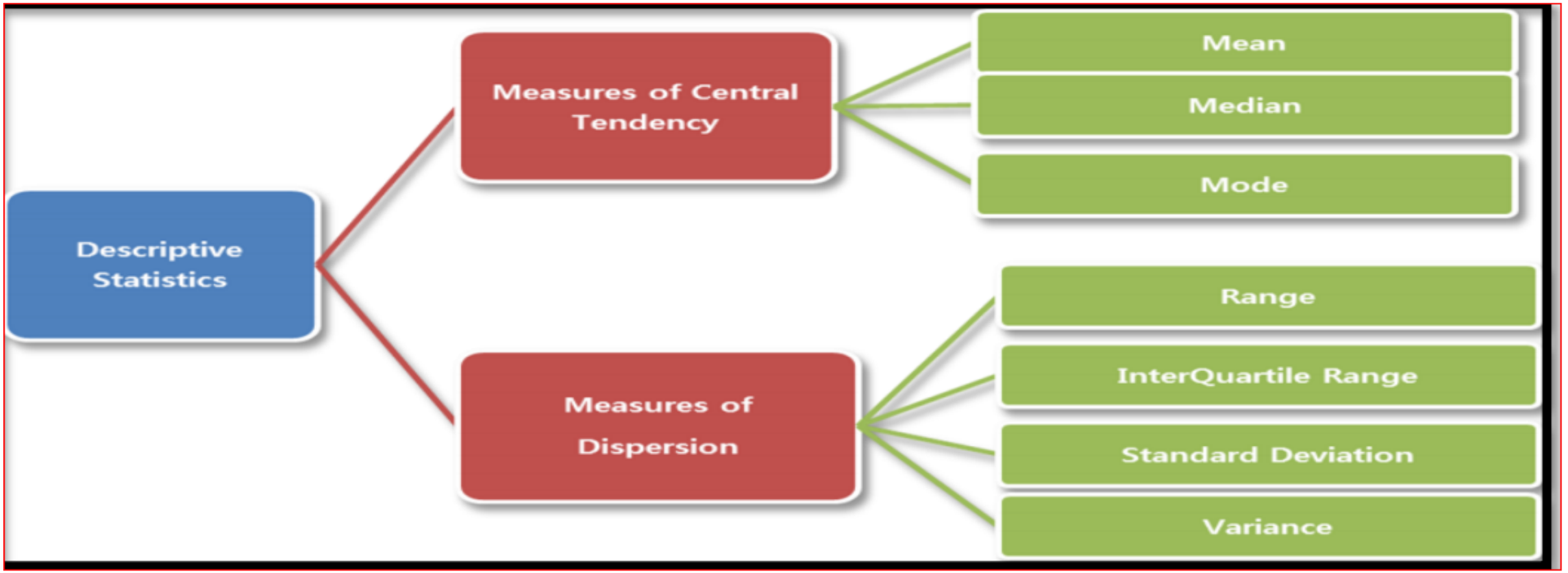
What’s cool about Pandas is that it takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to excel / SPSS

Why to use Pandas?

Wide range of functionalities such as

- Reading different varieties of data
- Functions for filtering, selecting and manipulating data
- Plotting data for visualization and exploration purpose

Understanding data using Pandas



..\Codes\Pandad and Undersytanding data using descrptives

Matplotlib

- Matplotlib can be used for creating plots and charts.
- The library is generally used as follows:
  - Call a plotting function with some data (e.g. .plot())
  - Call many functions to setup the properties of the plot (e.g. labels and colors).
  - Make the plot visible (e.g. .show())

Understanding data using Visz

..\Codes\Introduction to matplotlab lib and understand data using viszvilizing

Scipy - Inferential Statistics

Feature / Response	Continuous	Categorical
Continuous	Person’s correlation	LDA
Categorical	ANOVA	Chi-square

Parametric test

- ..\Codes\Inferential statiscs -Parametric test

`

Non Parametric test

- ..\Codes\Inferential statis 2 - Non paramertic tests

Broadly, there are 3 types of Machine Learning Algorithms

Supervised Learning:

This algorithm consist of atarget / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data.

Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

Unsupervised Learning:

In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention.

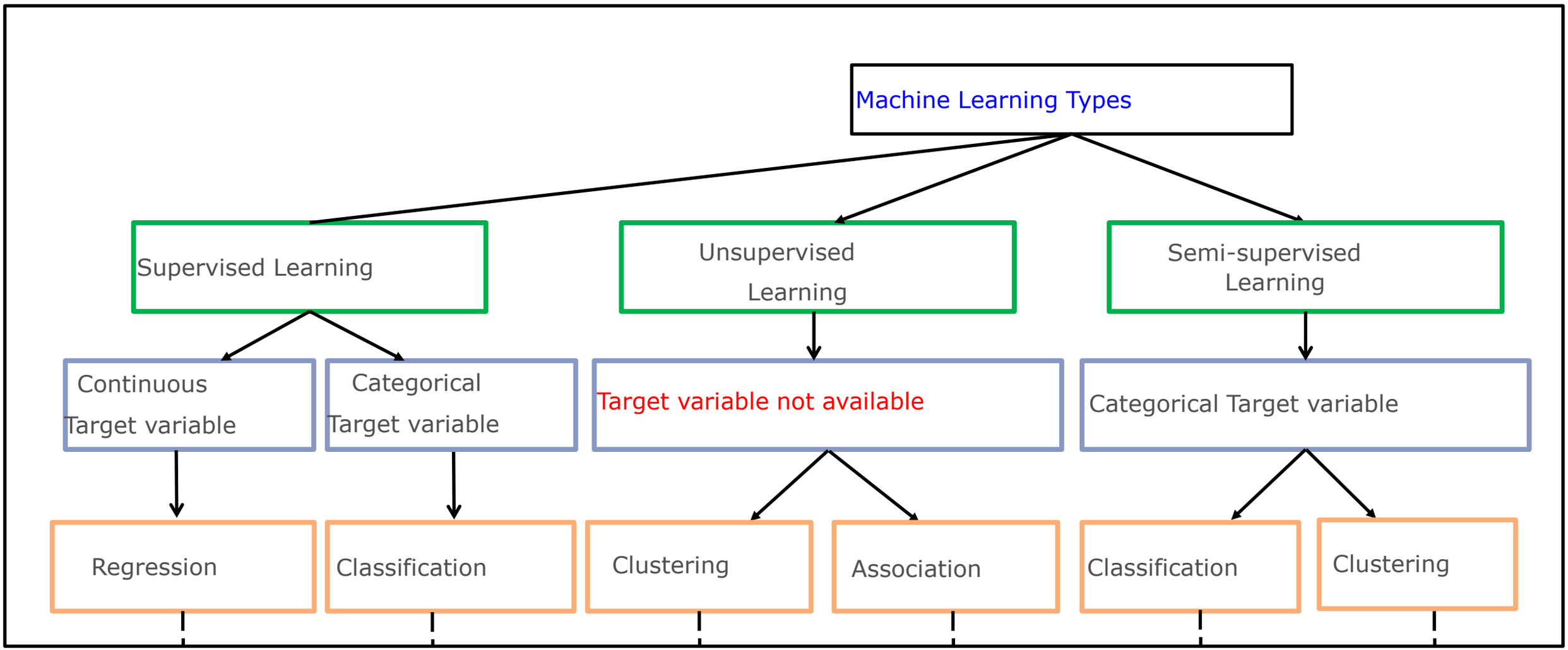
Examples of Unsupervised Learning: Apriori algorithm, K-means.

Reinforcement Learning:

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.

Example of Reinforcement Learning: Markov Decision Process

Machine Learning Algorithms



Linear Regression

Linear Regression is used when we want to predict an outcome variable that is interval / continuous with a set of predictors that are also interval / continuous. While categorical / nominal data can also be included,

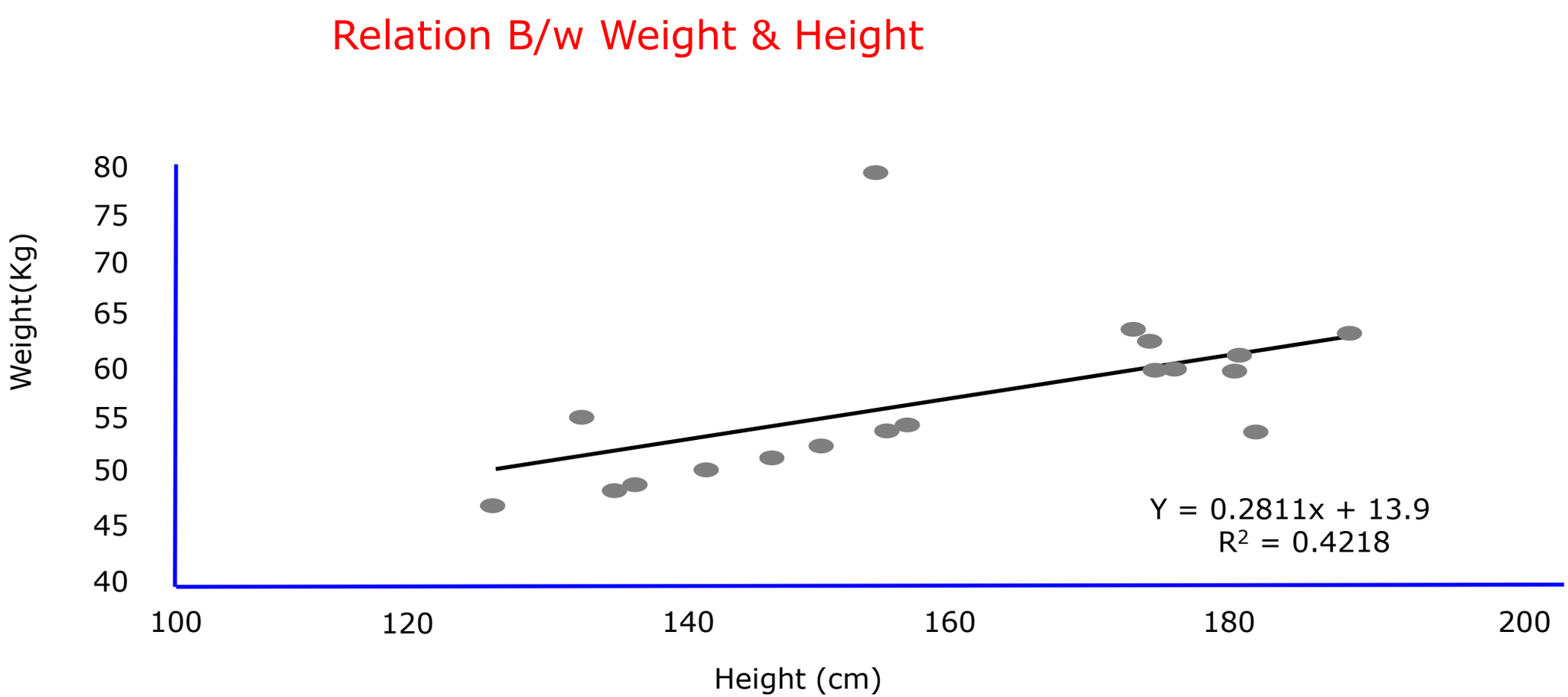
The representation of linear regression is an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), byfinding specific weightings for the input variables called coefficients (B)

For example: y = B0 + B1 * x → Simple Regression

(Y = B1*X1+ B2*X2+ B3*X3+ . . . + A -→ Multiple Regression

These coefficients B0 and B1 are derived based on minimizing the sum of Squared difference of distance between data points and regression line.

Linear Regression Contd…



Linear Regression Contd…

Linear Regression is mainly of two types:

Simple Linear Regression

Simple Linear Regression is characterized by one independent variable.

Multiple Linear Regression

Multiple Linear Regression(as the name suggests) is characterized by Multiple(more than1) independent variables.

Some good rules of thumb when using this technique are to remove variables that are very similar (correlated) and to remove noise from your data, if possible.

It is a fast and simple technique and good first algorithm to try.

..\case study1

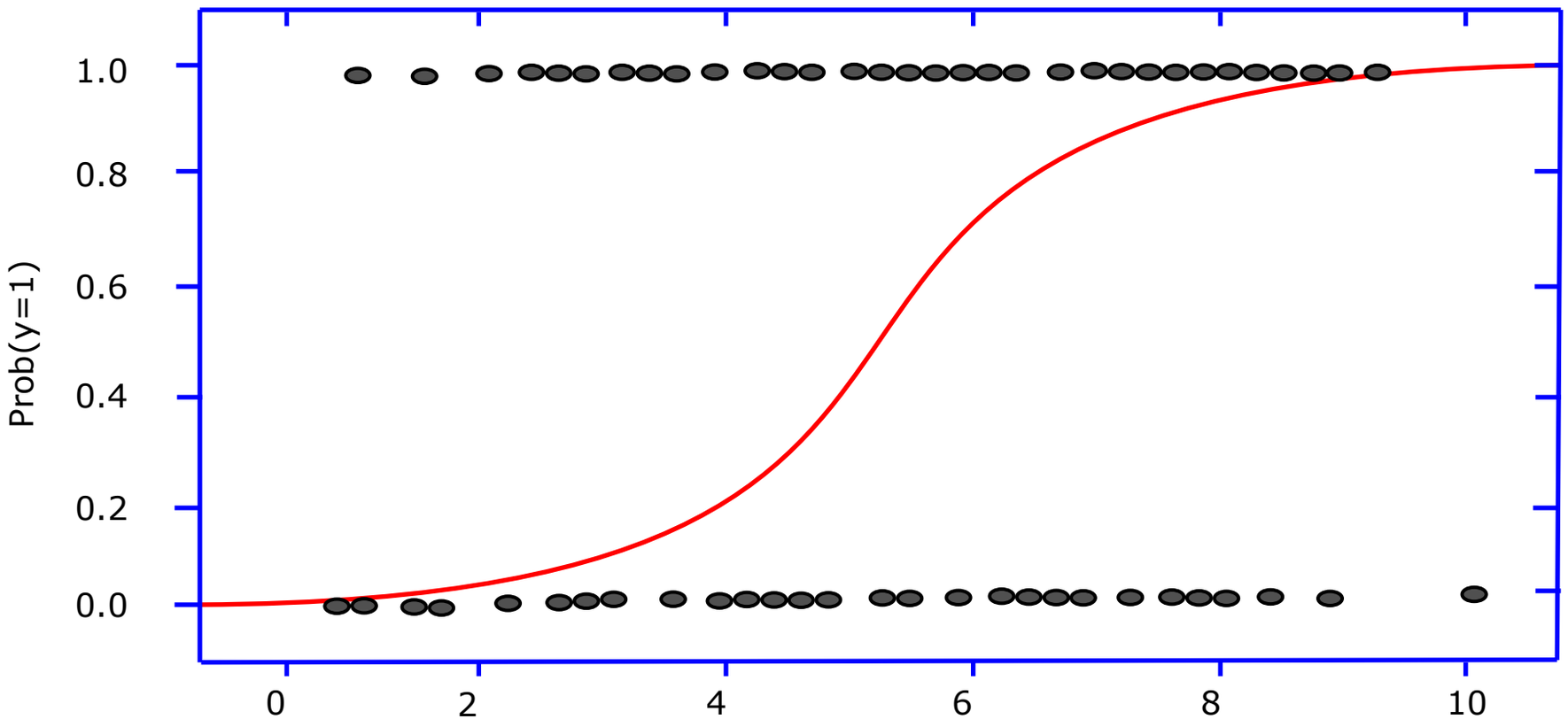
Logistic Regression

Logistic regression is a classification algorithm and it is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s).

In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression.

Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

Logistic Regression Contd…



Logistic Regression Contd…

Logistic regression does not directly model Y (dependent variable).

Logistic regression transforms the dependent into a logit variable (natural log of the odds of Y occurring or not occurring, which is ln(p/1-p)) and uses maximum likelihood estimation (MLE) to estimate the coefficients.

$$\text{prob}(\text{event}) = \frac{\exp^{(B_1 * X_1 + B_2 * X_2 + A)}}{\left(1 + \exp^{(B_1 * X_1 + B_2 * X_2 + A)}\right)}$$

Above, p is the probability of presence of the characteristic of interest. It chooses parameters that maximize the likelihood of observing the sample values (maximum likelihood estimation) rather than that minimizes the sum of squared errors (like in ordinary regression).

Like linear regression, logistic regression does work better when you remove attributes that are unrelated to the output variable as well as attributes that are very similar (correlated) to each other. It's a fast model to learn and effective on binary classification problems.

..\Case study2