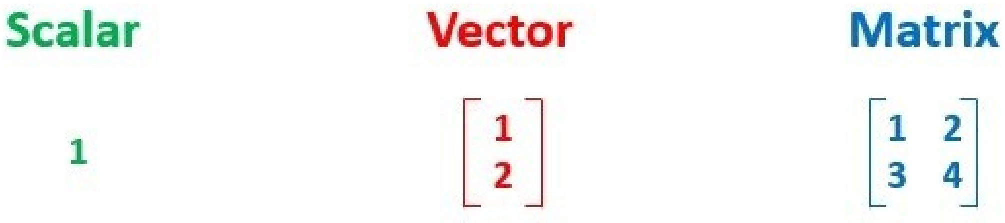
UNIT II	**Linear Algebra & Statistical Modeling for Data Science**	06 - Hrs
	Linear Algebra for Data science, Solving Linear Equations, Linear Algebra - Distance, hyperplanes and half spaces, Eigen values, Eigenvectors, Statistical Modeling, Random Variables and Probability Mass/Density Functions, Sample Statistics, descriptive statistics,notion of probability, distributions,Mean, variance, Covariance, Hypotheses Testing, Type 1 and Type 2 errors. Testing for parameters of a normal distribution and for percentages based on a single sample and based on two samples. Introduction to the chisquared test. The concept of p-value. Mean-square estimation and Kalman filtering.	

What is Data Science?

- Data science is an amalgamation of different scientific methods, algorithms and systems which enable us to gain insights and derive knowledge from data in various forms.
- Various organizations like Google, Facebook, Uber, Netflix, etc. are already leveraging data science to provide better experiences to their end users.
- Although data science techniques have been conceptualized and in use for several decades now, the current demand for data science is fueled by the high availability of digital data, and resources for computation.

Linear Algebra Concept

- Linear Algebra is a branch of mathematics. It provides basic structures to represent data and various numerical methods and tools to solve problems.
- The basic building blocks of Linear Algebra are scalar, vector, and matrix. A Scalar is a quantity, described by a numeric value. A Vector is an ordered collection of scalars. A Matrix is a collection of vectors.



Scalar

1

Vector

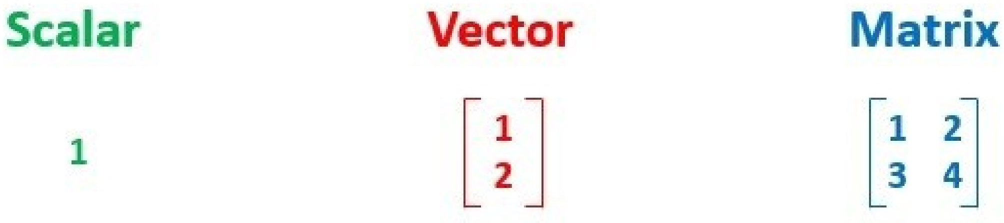
$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

Scalar Means

- Scalar is a single value – an integer, a boolean, or perhaps a string – while a compound is made up of multiple scalars (and possibly references to other compounds as well). When a distinction is needed between single/simple/atomic values and compound values, “scalar” is used.


Scalar
Vector
Matrix
1
$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$
$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

Why Linear Algebra?

- Linear Algebra is a branch of mathematics. It provides basic structures to represent data and various numerical methods and tools to solve problems.
- The basic building blocks of Linear Algebra are scalar, vector, and matrix. A Scalar is a quantity, described by a numeric value. A Vector is an ordered collection of scalars. A Matrix is a collection of vectors.

Why Linear Algebra for Data Science??

- Data Science deals with extracting meaningful insights from the data. We need to represent data and perform numerous operations on them to extract insights. Linear Algebra helps in representation and operations on data in Data Science and Machine Learning.
- **Let’s understand the application of Linear algebra in Data Science with the following examples.**
- **Deep Learning:** Representing input to the model and model parameters as vector and matrices and making calculations using Linear Algebraic operations.
- **Image data:** Representing images as matrices and doing various geometrical transformations on them using Linear Algebraic operations.
- **Recommender system:** using linear algebraic concepts to measure similarity.

Linear Algebra Applications

- **Multiple linear regression:** Solving multiple linear equations using linear algebra
- **Feature extraction (PCA):** using the linear algebraic concept of eigenvalue and eigenvectors
- **One hot encoding:** A matrix representation of the encoding

Linear Algebra With Example

- Let us explore an example to understand these building blocks.
- Consider a student who has taken 3 exams for a given subject. The **scores** are **80%, 80%, and 85%** respectively. Based on the difficulty level, the professor has allocated **weights of 30%, 30%, and 40%** to the exams respectively.

  Problem
- Calculate the weighted mean of the marks obtained by the student.
- The weighted mean **'W'** is calculated by the formula:

$$W = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

where **'wi'** represent the weights and 'Xi' represents the corresponding marks.

Linear Algebra With Example

- Here, individual score and weightage are considered as scalars. The collection of the scores and the corresponding weights can be represented as vectors as shown below:

  x = [80, 80, 85]

  w = [0.3, 0.3, 0.4]

- In Linear Algebra, an operation is defined called ‘dot product’ which helps in multiplying two vectors. The dot product of the vectors w and X is:

$$\mathrm{w}\ .\ \mathrm{X} = w_1 X_1 + w_2 X_2 + w_3 X_3$$

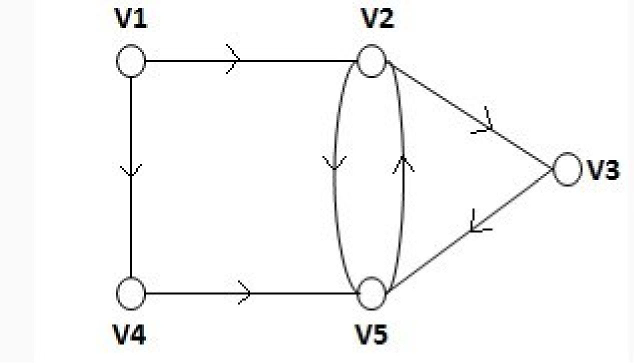
Linear Algebra With Example

$$\mathrm{w}\ .\ \mathrm{X} = w_1 X_1 + w_2 X_2 + w_3 X_3$$

- This dot product will result in finding the weighted mean of the marks (which is a scalar).
- Linear Algebraic operations are applied to the entire object (vector/matrix) instead of individual data points one at a time. This technique is called as Vectorization and it is more efficient while dealing with operations on large data.

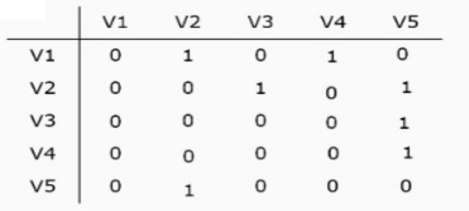
Example of Linear Algebra

- Let us explore another example where Linear Algebra is used for data representation.
- Consider a social networking site where five visitors are linked with each other as depicted in the graph below:



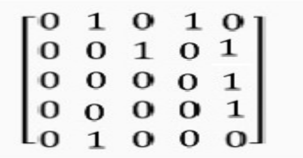
Problem-Linear Algebra

- How to use these relationships to extract more information about them?
- These relationships can be converted into a relationship matrix in which **'1'** indicates related and **'0'** indicates not related as shown below:



	V1	V2	V3	V4	V5
V1	0	1	0	1	0
V2	0	0	1	0	1
V3	0	0	0	0	1
V4	0	0	0	0	1
V5	0	1	0	0	0

- From the above relationships, you can create a matrix for the directed graph as below:

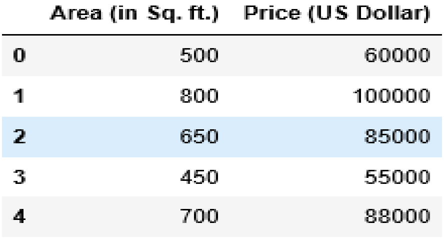


$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

- This matrix can be used as a data structure for representing graphs in computer programs for further computation.

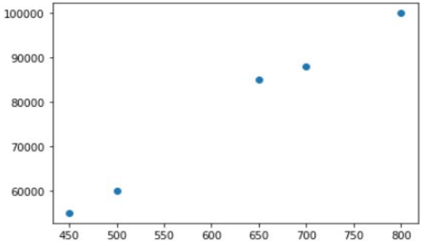
Introduction – Linear Equations

- Consider a scenario where you need to predict the price of a house given its area in square feet. The prediction can be made based on historical data of the houses sold in that region as given below:



	Area (in Sq. ft.)	Price (US Dollar)
0	500	60000
1	800	100000
2	650	85000
3	450	55000
4	700	88000

- Predicting the price of a house can be achieved by finding the relationship between Area and Price from the given data.
- The following is a scatter plot of Area and Price:



Introduction – Linear Equations

- This plot indicates that there is a linear relationship between the variables Area and Price. And the linear relationship is represented by an expression as shown below:

  Y = β0 + β1x1 + ϵ ------- (1)

  where,

  Y is the target variable ‘Price’

  x1 is the variable ‘Area’

  β0, β1 are coefficients

  ϵ is the random error
- When there are more independent variables in the dataset, the target variable can be in general represented as: Y = β0 + β1x1 +…+ βrxr + ϵ
- To find the optimal values of the coefficients β0 and β1, you can proceed as follows:
- 1. Find the sum of squares of the errors as given below:

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

- In the above expression, A and B are estimators of the coefficients β0 and β1.

Solving a System of Linear Equations

- Solving a System of Linear Equations implies finding the value of the variables such that each of the equations is satisfied.
- Consider a System of Linear Equations as shown below:

$$\begin{aligned} x_1 + 3x_2 - x_3 &= 4 \\ 2x_1 + 5x_2 + 4x_3 &= 19 \\ 2x_1 + 3x_2 - x_3 &= 7 \end{aligned}$$

- This System of Linear Equations can be represented in the form Ax = b as:

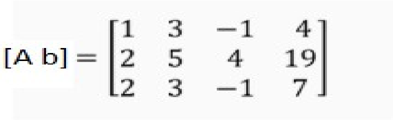
$$\begin{bmatrix} 1 & 3 & -1 \\ 2 & 5 & 4 \\ 2 & 3 & -1 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 19 \\ 7 \end{bmatrix}$$

- Where,

$$A = \begin{bmatrix} 1 & 3 & -1 \\ 2 & 5 & 4 \\ 2 & 3 & -1 \end{bmatrix}, b = \begin{bmatrix} 4 \\ 19 \\ 7 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

- In general, the matrix A is of dimension (m, n), x is a column vector of dimension n and b is a column vector of dimension m. A is called the Coefficient matrix and the matrix [A b] is called the Augmented matrix.

Solving a System of Linear Equations



$$[A\ b] = \begin{bmatrix} 1 & 3 & -1 & 4 \\ 2 & 5 & 4 & 19 \\ 2 & 3 & -1 & 7 \end{bmatrix}$$

- In general, the matrix A is of dimension (m, n), x is a column vector of dimension n and b is a column vector of dimension m. A is called the Coefficient matrix and the matrix [A b] is called the Augmented matrix.
- Note: For a given coefficient matrix A of dimension (m, n), if m = n and A is a Full Rank matrix, then, there exists precisely one solution for the System of Linear Equations.

- In the previous example of the System of Linear Equations, dimension of A is (3, 3) and its rank is 3. Therefore, there exists precisely one solution for the Linear Equations.