**Data science project**

1. Business problem

2. Data acquisition

* + 1. Logs
    2. web servers
    3. databases
    4. API’s
    5. Online Repositories

3. Data Preparation

1. Data cleaning

Inconsistent datatypes, misspelled attributes, missing and duplicate values

1. Transformation

Use Talent & informatica tools etc.

4. Exploratory Data Analysis

Defines and refines the selection of feature variables that will be used in the model development

most important step – because if didn’t to it correctly we will create inaccurate model

5. Data Modeling

Core activity of the data science project

Apply machine learning techniques repository like KNN, Decision tree, Naïve Bayes to identify the model that best fits the business requirement

Then trains the models on the training data sets and test to select the best performing model (For modeling the data we can use python, R or SAS)

6. Visualization and Communication

Tableau, power BI, like tools can use

7. Deploys & maintains

**What is Data Science?**

Data science is the field of study that deals with modern scientific techniques, statistical methods, and algorithms to derive insights from vast volumes of data

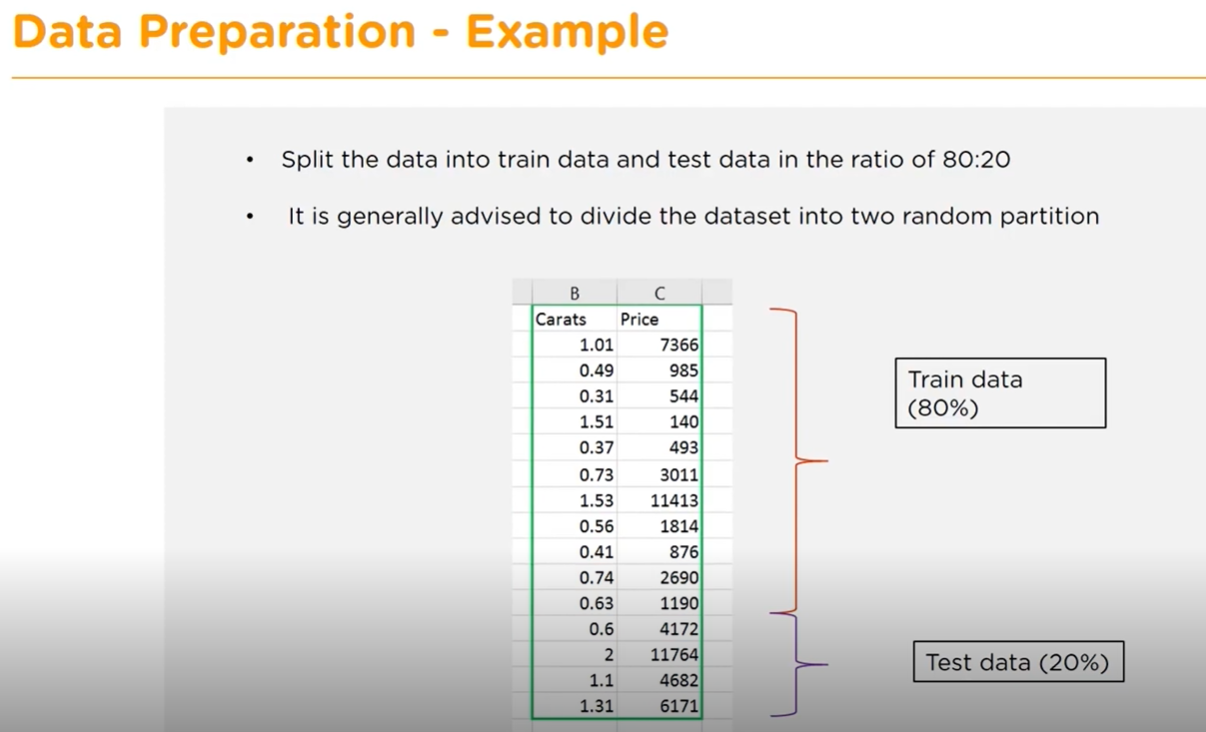
**Need for Data Science**

* Find unseen patterns within data
* Analyze and draw insights from the data
* Solve business problems and make decisions

**Data science lifecycle**

* Data discovery
* Data preparation – most time consuming

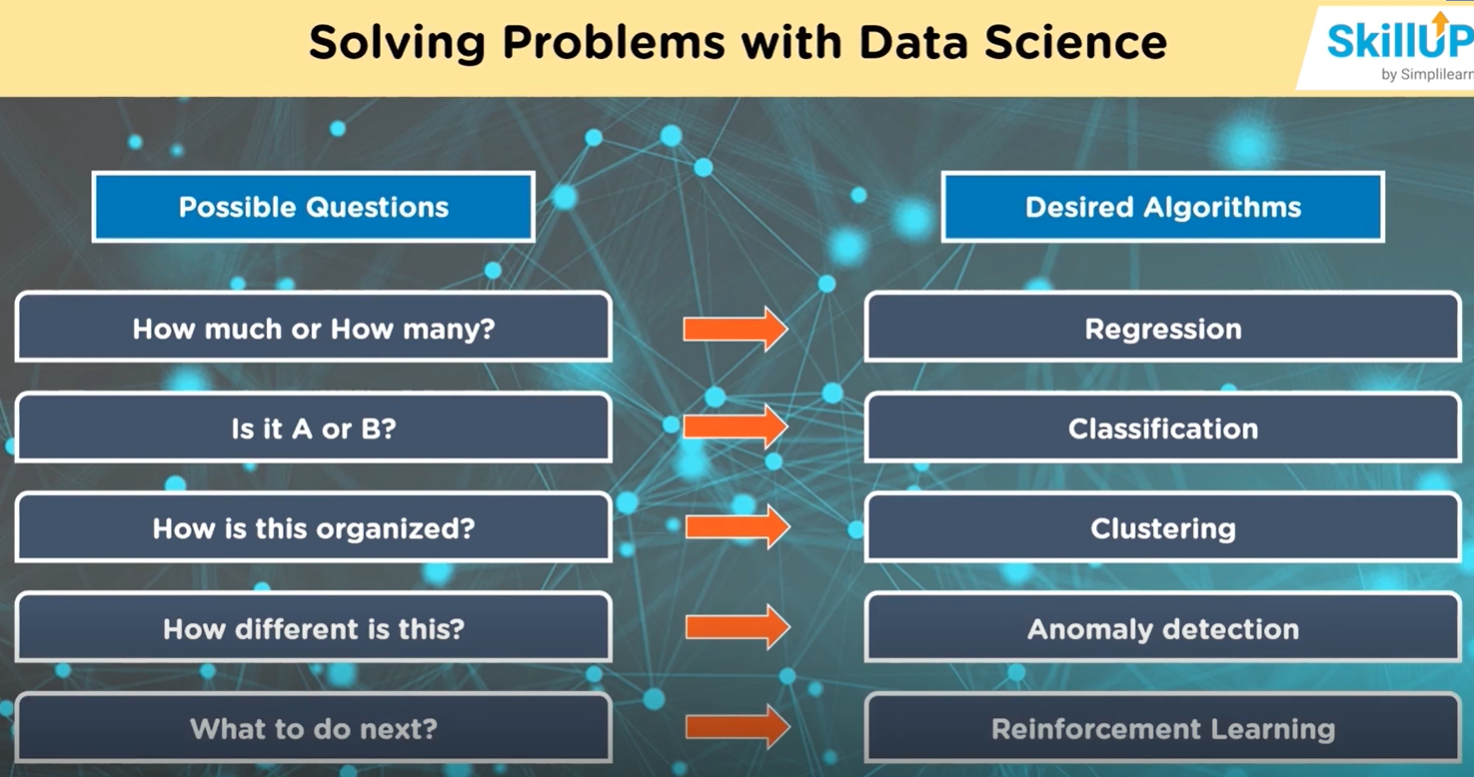
Ex:

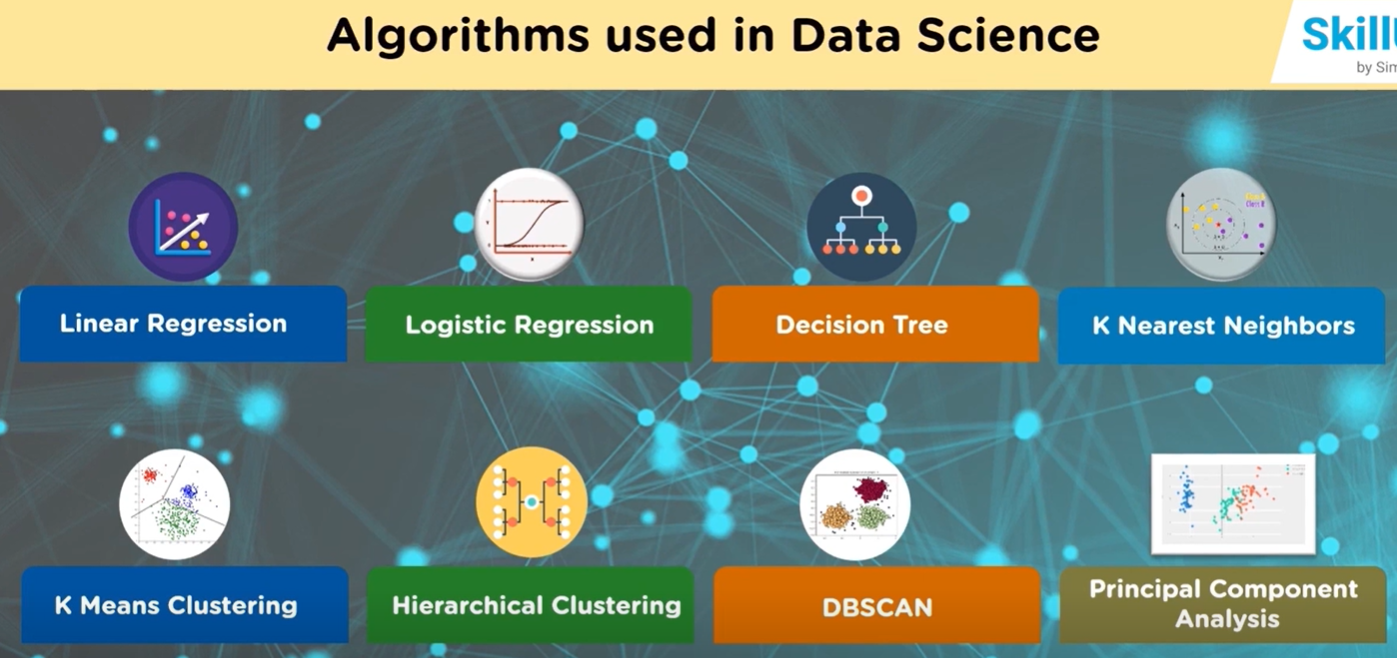


* Exploratory data analysis
* Data modeling
* Interpret the results

**Data science components**

* Mathematics
* Statistics
* Domain expertise
* Data engineering
* Data Visualization
* Machin learning





**3 Assential traits of a Data Scientist**

* Curiosity
* Common Sense
* Communication skills

**Prerequisites for Data Science**

* Machine learning
* Mathematical Modeling
* Statistics
* Computer Programming
* Understanding of Databases

**Tools and Skills used in Data science**

Data Analisis

Skills – Python, R

Tools – SAS, Jupyter, R studio, MATLAB, Excel

Data warehousing

Skills – ETL, SQL, Hadoop, Apache spark

Tools – Informatica, Talend, AWS

Data Visualization

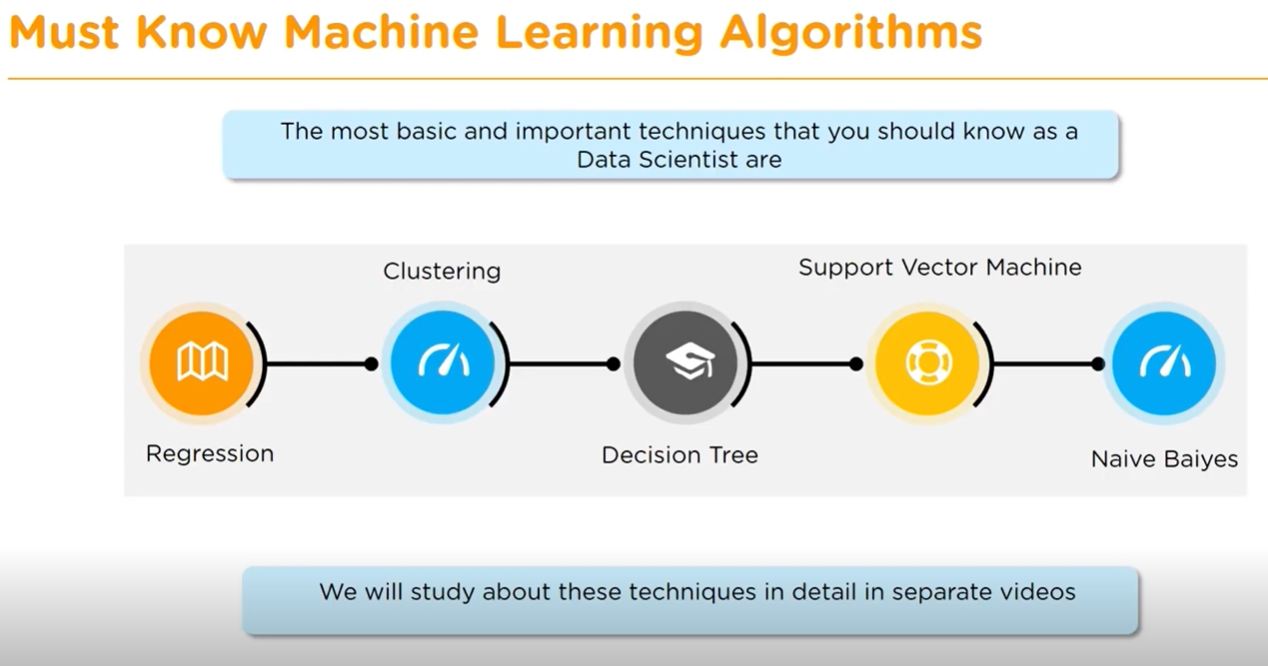
Skills – R, Python

Tools – Tableau

Machine Learning

Skills – python, algebra, statistics, ML algorithms

Tools – Spart Mlib, Azure ML studio



**Top python libraries for Data Science**

* Tensor Flow

For high performance numerical computation. Data in TensorFlow are represented as tensors, which are multidimensional arrays. Better visualization graph visualizations.

* Numpy

For numerical python.

* Scipy

For scientific python

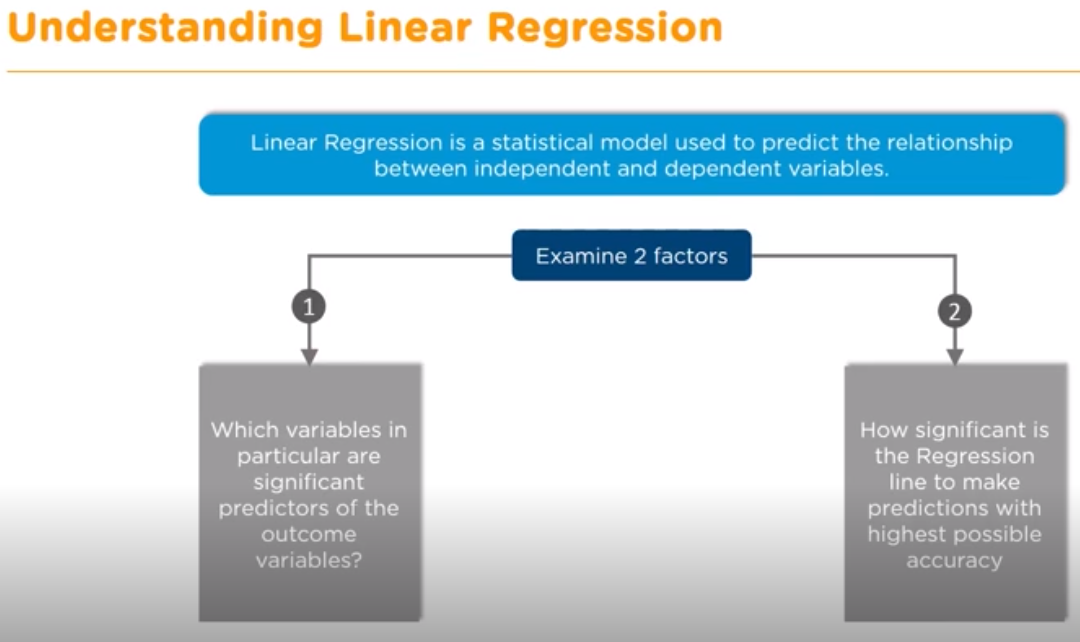
* PANDAS

Stands for python data analysis library

* Matplotlib

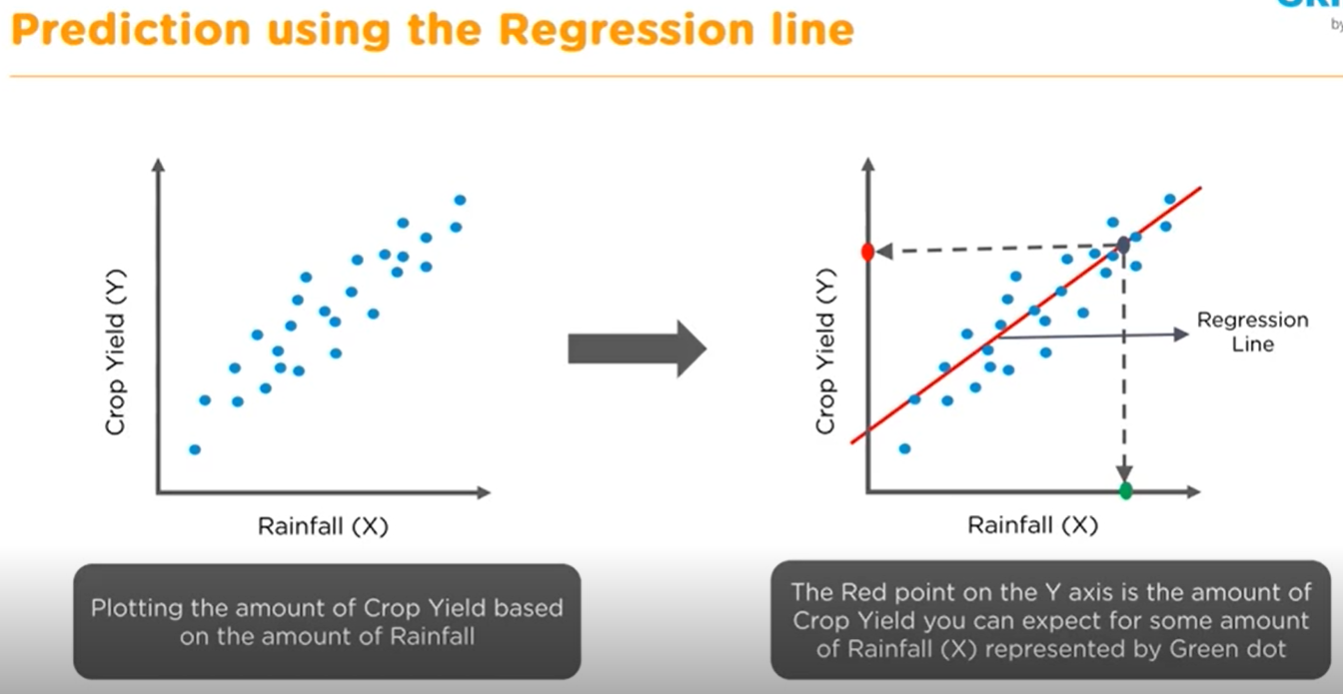
Ploting library of python

**Linear Regression**



Graphical user interface

Description automatically generated with low confidence



Chart

Description automatically generated

A picture containing table

Description automatically generated

A picture containing calendar

Description automatically generated

A picture containing table

Description automatically generated

Chart, scatter chart

Description automatically generated

**Multiple Linear Regressions**

Diagram

Description automatically generated

**Logistic Regression**

