Data Science is generation of actionable knowledge directly from huge amount of complex data. The goal of data science is to gain insights and knowledge from any type of data — both structured and unstructured

Structured data:

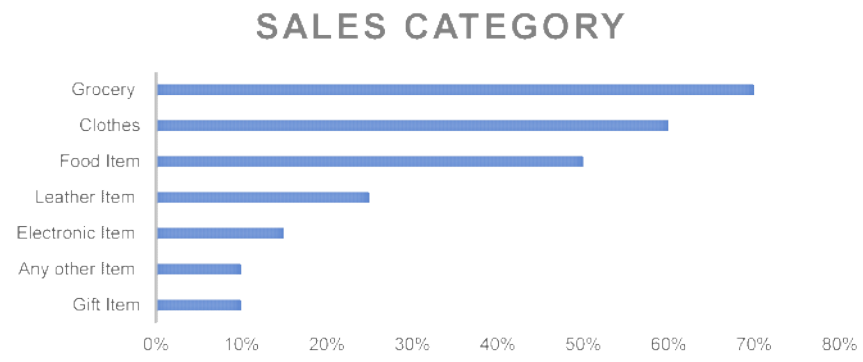
Rows and columns

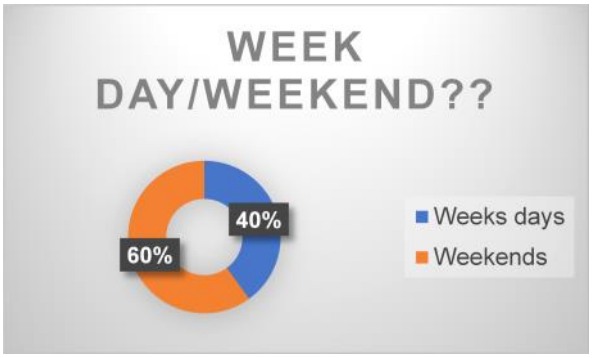
Name , age, height, weight, location, education, working, exp, tech skills
1
2
3
4
5

Banking, hospitals, universities, sales, insurance, telecom, pharamacy, tours and travels, bookmy show,

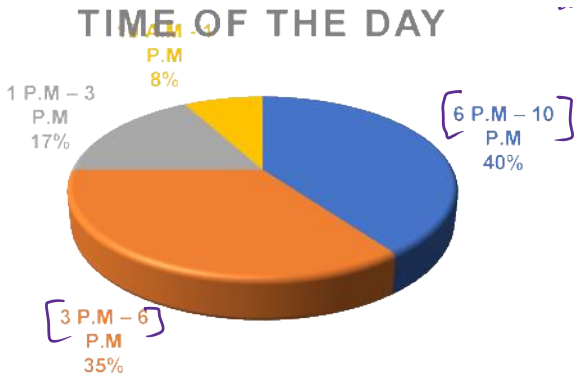
Unstructured data:

Any text data, google search, images, videos, comments, reviews, blogs, websites , feedback, whatstapp chat, scanned Reports, doctor's prescription,

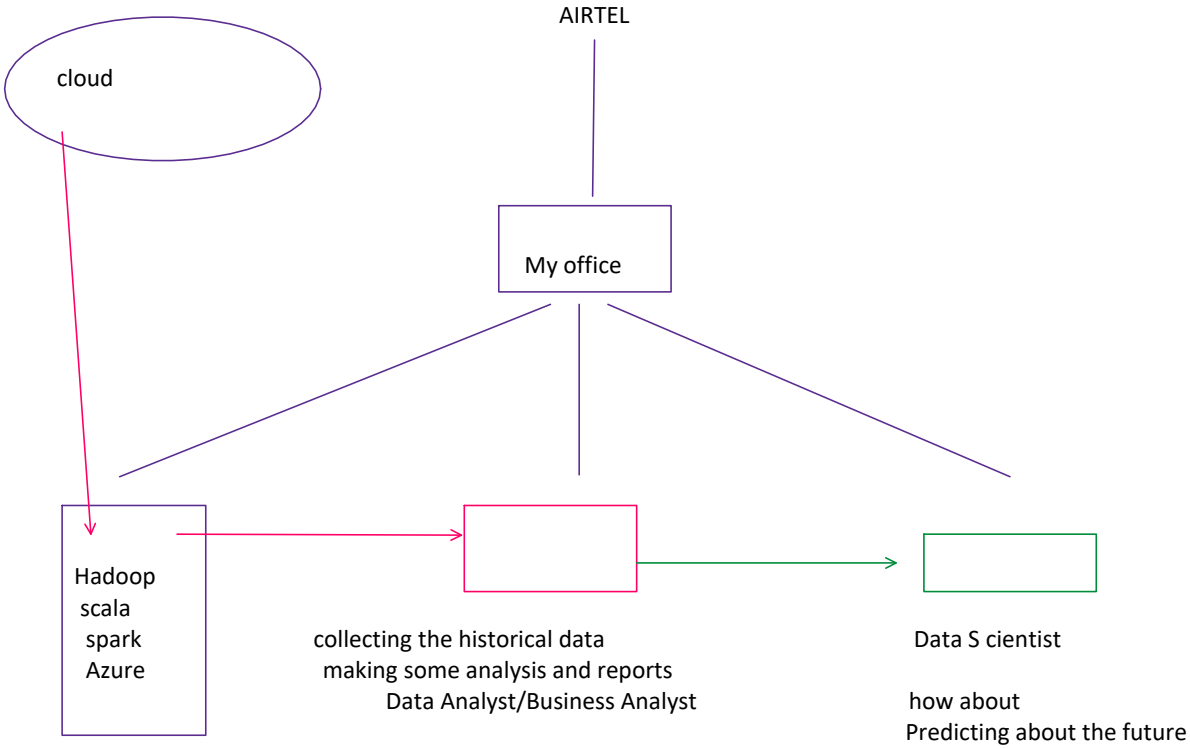




12:00 → 10%



==========



Big data engineer

Tableau, Power Bi, Excel, Python

Y = mx + C

==========

Basis statistics --> 2 days

Python --> scratch --> 2 days

Statistics + python --> 1 days

Probability , Test of hypothesis --> 10-12 hours

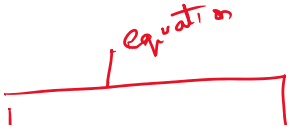
==========

Machine Learning:

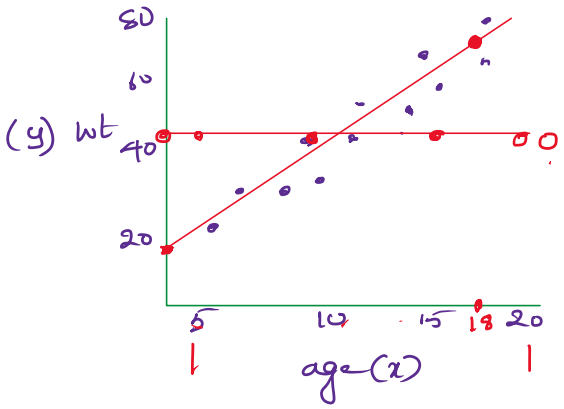
2 months -->

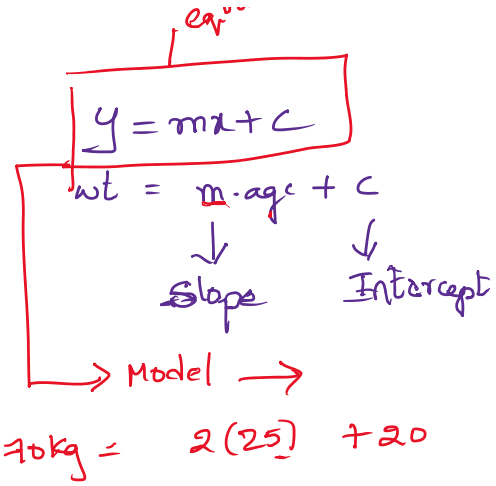
NLP --> 1 weekend

Deep learning --> 1 week end



Deep learning --> 1 week end





Value added courses:

AI classes --> 3 weeks
Hadoop , spark
Azure
Tableau
Powerbi
Core python
SQL --> 10 hrs

Assignments

20 --> assignment team

---> Project

==========

Steps behind the project life cycle of data scientist:

1. Framing the problem:

2. SQL -> collect the data from the server to our local system

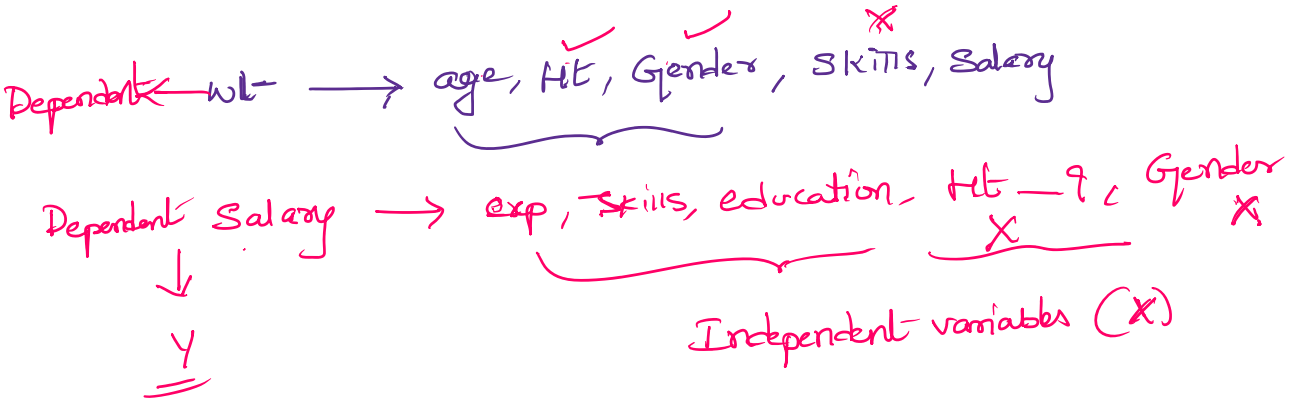
   SQL to Python

3. **Exploratory data analysis**:

   Structured data ---> EDA

   Unstructured --> preprocessing

Which columns are really important to us.



Scatter plot, box plot, histogram , bar graph

Data cleaning:

Data cleaning:

Boxplot

Data Transformation:

What is **transformation?**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

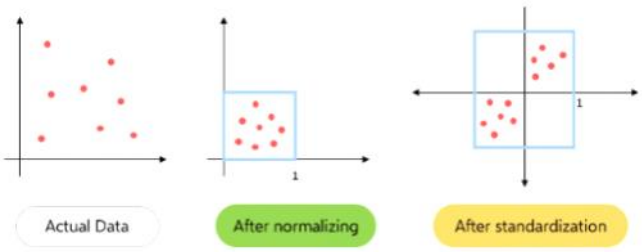
wt = 10 + 26 + 170

206 kgs = 206 yrs / cms

wt → Age + ht

26 yrs, 170 cms

units will be removed



Data partition:

Investment → Test ?

1600
① Training → 600 → model
② Test → 400 → tested

Selection of model:

Machine learning

$f(x) = x^2$ $\quad y = x^2$

$f(x) = 2x$

x → y

1, 2, 3 → 1, 4, 9

Bread (y) — x^2

$f(x) = 2x$

$f(x) = 2x+1$

Based of 'Y' x^2 / $2x$ / $2x+1$

1. Linear Regression → 2x+1
2. Logistic Regression
3. Support vector machine
4. Decision Trees
5. KNN classifier
6. Naïve Bayes classifier

Probability, matrices, distance method

Cross validation:

Model should be verified with various methods --> wrong results

Evaluation:

Deployment:

==========