Rainbow

Data Science and Big Data analytics PAGE: DATE: 1/06/2022

* Data Science : It is the field that helps in extracting meaningful insights from data using programming skills, domain knowledge and mathematical and statistical knowledge.
  > It helps to analyze the raw data and find the hidden pattern
  > It includes collecting data, analysing the data and building modles from that data.

* Big data:
  > It is collection of data that is huge in volume and yet growing exponentially with time.
  > It is a data with so large size and complexity that none of traditional data management tools can store it or process it effeciently.
  > It is a data of huge size.

* Appln of data science:
  1) In Search engines.
  2) In transport (Like, driverless cars)
  3) In finance. (To being safe from fraud and risk of loss)
  4) In E-commerce (To give recommendations and inhancing user experience)
  5) In Health care [Detecting and dignosing deasese on time, doing drug discoveries)

* Data Explosion:
  The rapid or exponential increase in the amount of data this is generated and stored in the computing systems, that reaches level where data management

1 Exabyte : 1 billion GB

Rainbow
PAGE:
DATE: / /

becomes difficult, it is called as data explosion.

* The key factors of data growth
1) Increase in storage capacity
2) Cheaper storage.
3) Increase in data processing capabilities.
4) Data generation and it get easily avalible.

* 5 V's of data science :
1) Volume → Amount of data generated
2) Velocity → speed of data generation
3) Variety → Types / formats of data which is generating
4) Veracity → Truthfulness of data
5) Value → Impact of data

Understand using medical industry example :

Volume	Velocity
Hospital and clinic generates huge amount of data. about 2,314 Exabyte of data generated annually.	Data get generated at very high speed In the form of patient records & test results

Variety	Refers to the variety of data like structured (excel), semi-structured (log files), unstructured data (x-ray files.

Rainbow

PAGE:
DATE: / /

Veracity (Truthfulness)

The accuracy and trustworthyness of generated data.

Value

Aanlysing the data gives benefits.
Here it will better disease detection, better treatment, Reduced cost

* Relationship between data science & Information sci.

* Information science :-
It includes storing and retrieving information.

* Data Science :
It is the discovery of knowledge or actionable Information in data.

* Data science and Information science are distinct but complimentary disciplines.

* Data Science is heavy on computer science and math's.

* Information science is more concentrated with areas such as library science, cognitive science & communications.

* Data science is used in business functions such as Strategy formation, decision making and operational processes.

Rainbow
PAGE:
DATE: / /

* Information science is used in areas such as knowledge management, data management and interaction design.

* Data science and Business intelligence.

Factor	Data Science	Business Intelligence.
Concept	It is field that uses maths, stats, tools to discover the hidden patterns in the data.	It is set of technologies, appl^n & processes that are used by the enterprises for business data analysis.
Focus	It focuses on future	It focuses the past and present.
Flexibity	More flexible as data resources can be added as per req.	Less flexible as its data sources need to be preplaned.
Method	Scientific method	Analytic method.
Complexity	Higher complexity compair to BI	Less complex/simpler compair to DS.
Expertise	Data scientist	Business user.
Questions	what will happen & what if	what happened.

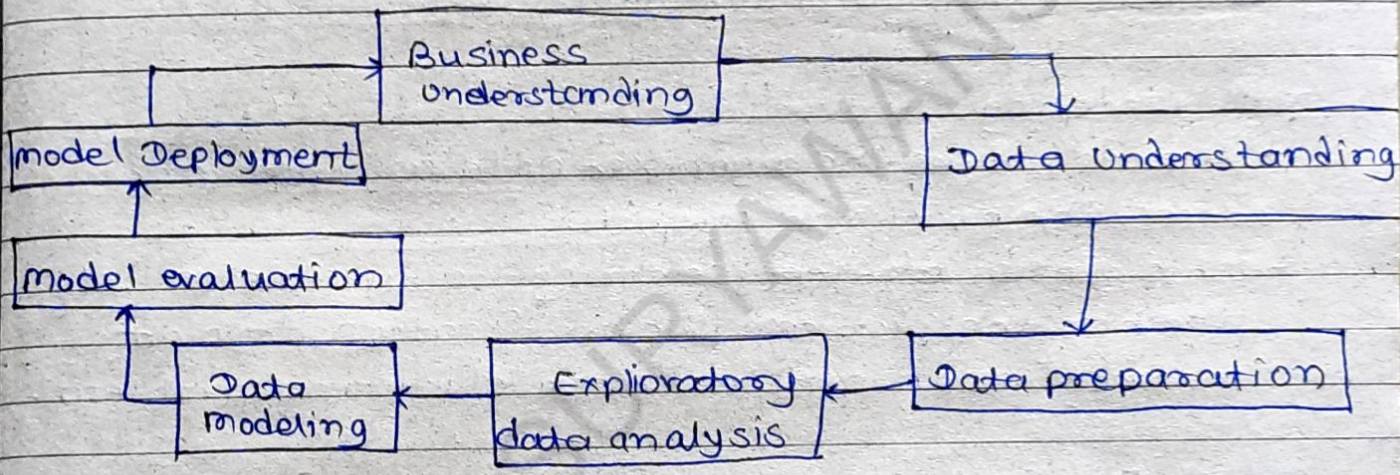
Rainbow

PAGE:
DATE: / /

Tools	SAS, MATLAB, BigML, Excel, etc.	Insight Squared sales Analytis, Klipfolio, cyfe, etc.

Data Science Lifecycle

It revolves around the use of machine learning and different analytical strategies



> Business understanding

The complete cycle revolves around the Goal bec this is the ultimate aim of analysis.

> Data understanding

This includes describing the data, their structure, their ~~or~~ revelence, their records type

> preparation of Data

This includes steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values

through either eliminating them or imputing them, treating inaccurate data through eliminating them

> Exploratory Data Analysis

This step includes ~~the~~ getting some concept about the answer and elements affecting it, earlier than constructing final model.

> Data Modeling

- A model takes the organized data as input and gives the preffered output.
- This consist of ~~steps~~ selecting the suitable model, whether the problem is classification problem, or a regression problem or a clustring problem.

> Model Evaluation

- Here the model is evaluated ~~for~~ checking if it is ready to deploy.
- The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics.
- we additionally need to make positive that the model confirms to reality.
- If we do not meet the required results we have to re-iterate the complete modeling process. Till the preffered stage of matrics is achieved.

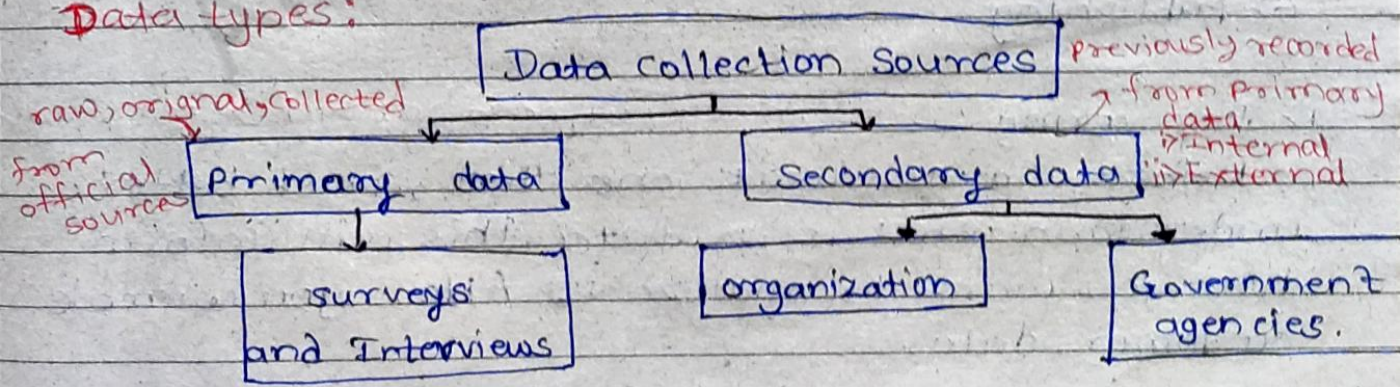
Rainbow
PAGE:
DATE: / /

> Model Deployment
- The model after a rigorous assessment is at the end deployed in the preferred structure and channel.

What is Data?
> Data is a different types of information that usually is formed in perticular manner.
> The quantities, characters or symbols on which operations are performed by a computer which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical or mechanical recording media.

Data collection
It means collecting information about something with an objective to analyze it or extract some meaningful information from it.
Ex. A company collecting info about product sales by their outlets.

Data types:



Rainbow
PAGE:
DATE: / /

⊙ Data wrangling

> Data wrangling is the process of cleaning, structuring and enriching raw data into desired data format for better decision making in less time.

> As the data amount and sources of data are increasing rapidly it is important to the large amount of data need to be organized for analysis.

> This process typically includes manually converting / mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

⊙ There are six iterative steps of data wrangling process.

1> Discovering
2> structuring
3> Cleaning
4> Enriching
5> Validating
6> Publishing

1> Discovering : (Knowing about data on which going to work)

Before starting work one that, must better understand what is in data, this will inform how you want to analize it / How you wrangle customer data.

Rainbow
PAGE:
DATE: / /

2> Structuring [Giving structure to data for better understanding]
Organizing data in desired format because data can be of any size or shape. This will help in analysing the data.

3> Cleaning [Removing unwanted things from data]
> Removing unwanted things from data which may cause issue in further · cleaning things like null values, wrong type of values in wrong format.
> Saving data from getting corrupted.

4> Enriching [Improving data quality]
> Analysing and making strategy to enhancing the data by adding some additional data which will help in enriching data.
> This will help in decision making more properly

5> Validating [Helps to increase quality and security of data]
> This is used for improving data consistency quality and security of data.
> E.g. Like mobile number validation or cross checking length of them of no invalid number should be in data.

6> Publishing [Publising data for accessing by other available to]
>Here we have wrangled data so we have to publish it, this is done by user or any software.

> This data need/should be available for other people for analysing or doing operations on that.

Data issues

Issues in data cleaning

1> Lack of validation.
2> Data from different sources
3> Personal names
4> Locations.
5> Dates
6> Numbers
7> Currencies
8> Languages
9> other issues :- spell. mistakes, etc.

Data Cleaning Methods

Some methods which are used to clean data:

1> Histograms — used to find out which values are being used less frequently.
2> Conversation Tables — Data issues & parameters are already known.
3> Tools — vendors offers some solutions for data cleaning are IBM, SAS, oracle, etc.
4> Algorithms — To fix the data, spell checking, etc.
5> Manually — cleaned by hand/by human interface.

Data integration

It is preprocessing method that involves merging of data from different sources in order to form a data store like data wearhouse.

* Issues in data integration:

① Schema integration and object matching.

(A) Emp.Id	name
1	A
2	B

(B) Emp.no	name
3	C
4	D

These are same but at time of integration system will face issue to detect that these are the same

② Redundancy → unwanted attributes.
If DOB is collected then their is no need of Age collection because we can calculate it from age. So age becomes redundant data.

③ Detection and resolution of data values conflicts.

Company (A)
Price shown
¥ 7 Z

Company (B)
price shown
¥ $

Here if we want to change Z into $ and if we do it like ¥7$ this will cause massive disaster / Data conflic-tion.

preprocessing → Preparing for further processes

Rainbow
PAGE:
DATE: / /

> There are two major approaches of data integration

1> Tight coupling →
- Data wear house is treated as an info retrival component
- Data is combined from diff sources to one physical source by process ETL → Extraction, Transformation & Loading.

2> Loose Coupling
- Takes the query from the user and transform it in a way that the source database can understand.
- The data stays in actual source database.

* Data Reduction. Like we have 1TB data we will reduce volm of it but info will remain intact.
→ Preprocessing technigue that helps in obtaining reduced representation of dataset from the available data set.
→ Integrity / quality of the data should remain maintained even after reduction in volm of the data. from reduced data
→ It should produce same ~~red~~ reslut which comes on orignal data.

> Methods of data reduction

① Data cube Aggregation
→ It is a process in which info^n is gathered and expressed in a summary form.
Used for stastical analysis.

→ Dataset in smaller volm.

Year 2017

Half year	Sales
H1	500
H2	300

→ getting sum aggregate.

Appling data cube aggre.

Year	Sales
2017	800
2018	900

Year 2018

Half year	Sales
H1	300
H2	600

② Dimension Reduction
Removes redundant attributes.

③ Data Compression.
Reducing the file size using diff compression technique like Huffman Encoding & run-length Enc.]

i> Lossy compression
The compressed data may diff to orignal data but it is useful.
e.g. image compression.

ii> Lossless compression.
→ simple and minimal data reduction.
→ Algo used to restore orignal pricise data.

Rainbow
PAGE:
DATE: / /

④ Numerosity Reduction.
Data is replaced by estimated / alternative mainly mathematical model.

⑤ Discretization & concept Hierarchy operation.
Data is replaced by range of value or higher hierarchy / higher level.

⊙ Data transformation.
→ It is data preprocessing technique that transform or consolidate the data into alternate forms appropriate for mining.

* Involved processes
i> Smoothing: Removing the noise from data
[like binning, regression, clustring, etc.]

ii] Aggregation: Summary or Aggregate fun^n applied
Used in construction of data cube.

iii] Generalization: Low level concepts are replaced with higher level concepts.
Eg. street → city / country

⊙ Points to be considered while transforming the data:
1. Searching the data → right data search
2. Filtering the info → Like taking Last year info from past 5 yr.
3. Right skills →
4. Right Tools.

5. Intelligent amalgamation. → converting big data to smart data.
6. Analyzing the info
7. Right strategy.

iv] Normalization:

Attributes values are normalized by scaling their values so that they fall in specified range.

For -

2
40
100
500 ← max_x
1 ← min_x

Scaling & converting in 0-1 range → [0-1] → range

i> min-Max Normalization

$$v' = \frac{v - min_x}{max_x - min_x}$$

v' → new value
v → original attribute value
min_x → min value of attribute
max_x → max value of attribute

For 2 $v' = \frac{2-1}{500-1} = \frac{1}{499}$ } normalised value for (2)

ii> Z-Score normalization

→ zero mean

$$v' = \frac{v - \bar{x}}{\sigma_x}$$

v' → new value
v → orig. attribute value
$\bar{x}$ → mean of attribute
σ_x → Standard deviation

Rainbow
PAGE:
DATE: / /

Data Discretization or Binning

→ It divides the range of attributes into small intervals so as to reduce number of values for a given continuous attribute.

↳ Splitting : Top-down [Attribute is splitted into range of values.

↳ merging : Bottom-up [Initially we consider later remove some during merging

↳ Supervised : class info is known

↳ unsupervised : class info is not known.

* Noisy data : data with lots of meaningless info in it called noise.

* Data smoothing tech. :-

1> Binning → smoothing data by consulting values around it.

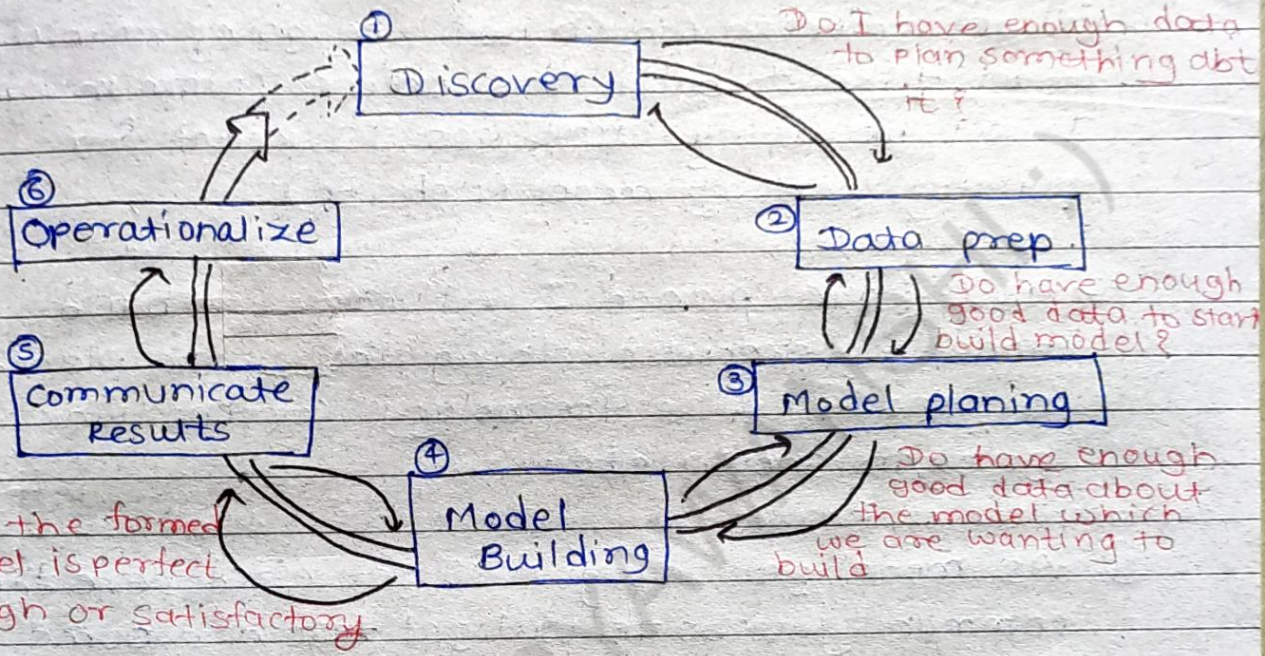
2> Regression → It confirms data values to a fun.

3> Outliner analysis → Outliners may be detected by clustering..

Unit 3

Big Data Analytics Life Cycle.

Rainbow
PAGE:
DATE: 2 / 6 / 2022



* Example of college community building.

Discovery

1> First need to discover how many students are intrested and need to gather their info like interest of field, name, etc and also who want to get stake in it.

2> Data preparation.

2> After getting/gathering data of students I will need to clean, condition and select the appropriate data from that gathered raw data. And prepare the required data.

model planing

3> After getting desired data I will start to plan model which further will build. In this similar data will get clustered, Associated with rules

or get classified. And methods, technique will get selected which will be going to built model. Like to build this online community I will be using HTML, CSS, JS, React for building working model (website)

Model building

4> As I have planed I will start to build model by executing thing those are planed for model: Like technology, content which going to shown etc.

Communicating Result

⑤ As model get ready it is get tested here the community website will get test and some discussion will be done with community members that what can be improved in this model / website.

Operationalizing

⑥ After making final changes our community website (model) is ready to get live for usage or in operation.

Phase-1

* Discovery phase

↳ completely understanding info about that business domain and business process. Recognizing key metrics and KPI's

↳ Evaluating the available resources and going through the process of framing the business problem as an analytic hypothesis.

Rainbow
PAGE:
DATE: / /

↳ Resources planning
  ↳ Data — From where, how much, how we will get data.
  ↳ People → Formation of teams, managing people, understanding what is going done by whom.
  ↳ Time → managing it and keep track of it Like deadlines etc.
↳ Framing the problem :- defining the problem and trying to find its solutions.

Phase 2 - Data Preparation.

↳ Analising raw data.
↳ obtaining, cleaning, aligning and examining the data.
↳ Transforming and enhancing the data.
↳ ETL (Extract, transform, load)
  - Extract data from sources
  - transform data in consistent format
  - Load data when ever needed
↳ Data conditioning (standarzation) — removing ununiformaty.
↳ Data selection — removing redudant data.

After this phase we have quality of data to move further.

Phase 3 - Model planing

↳ Methods used to build model

Clustering, Association rules, Regression & classification

Clustering	Association rules	Regression	Classification
Combining similar kind of data	If data objects share a bond anywhere then they can be associated.	keeping dependant and independent variables in rel^n	Labling data.

↳ Selection of variables and methods / techniques to build a model using avalible data.

Phase 4 - Model Building
↳ Building model according to planning and avalible data.
↳ Executing those techniques and methods
↳ Additional tools, requirements which are required.
↳ Manipulation of the data, calculating the reliability of the data, and determining the quality and predictive powers of the formed model.

Phase 5 - Communication Result.
↳ Success or failure
Discussing the analytic process and model was successful and accomphishes the required analytic goal of the project.
↳ Key finding, business value & summarized narration to stakeholders about result and model.

Pase 6 - Operationalize
↳ Providing the final suggestions, reports, meetings, code and technical documents.
↳ Final report & briefing.
↳ Risk managed.

Rainbow

PAGE:
DATE: / /

Case study - GINA (Global Innovation Network & Analysis)

↳ EMCs GINA is team of senior technologists placed in center of excellence (COE's) All over the world.

↳ Goal : connect employees all over the world to drive innovation, research & partnerships

↳ As per GINA team consideration it would offer an interface to share ideas globally and enhance sharing of knowledge between GINA members who are not at one place.

↳ A data repository is made/created to store structured and unstructured data to achieve three imp goals:

1> Store formal as well as informal data

2> Keep track of research from technologists all over the world

3> To enhance the operations and strategy, extract data for patterns and insights.

Unit III

Predictive Big Data Analytics with Python.

Rainbow
PAGE:
DATE: / /

- **Python** → created by Guido van Rossum
  → released in 1991
- **used for**
  - ↳ web development (server-side)
  - ↳ Software dev.
  - ↳ Mathematics
  - ↳ System Scripting
  - ↳ connect database systems. Read/modify files.
  - ↳ Rapid prototyping.
  - ↳ Handle big data and perform complex math.
- **Importance**
  - ↳ works on diff. platforms
  - ↳ simple syntax and this allow to write code in less lines.
  - ↳ Python runs on interpreter system. → can get executed as written and helps in fast prototyping.
  - ↳ Can treated in procedural way, an object oriented way.
- **Syntax**
  - ↳ designed for readability & has some similarities to English
  - ↳ use new line to complete command
  - ↳ Relies on indentation, using whitespace to define scope.

collection of functions and methods

⊙ Python Libraries for Data Processing and Modeling.

1. Pandas

↳ used for data analysis and data handling.
↳ It was created as community library project and released around 2008
↳ Provides easy to use and high performance data structures and operations for manipulating data.
↳ Multiple tools avalible for reading & writing data
↳ It can take data from files like csv, excel, etc or a sql data base as well as create a python obj known as data frame.

2. Numpy

↳ used for numerical computing on data in form of large arrays and multi-dimentional matrices
↳ Multidimentional arrays are the main objects in Numpy where their dimentions → axes
↳ number of axes → rank.

3. Scipy

↳ used for scientific computing and technical computing
↳ community library project and released around 2001
↳ scipy is built on the Numpy array object and it is part of Numpy stack which also includes other libraries and tools such as Matplotlib, Sympy, pandas etc.

Rainbow
PAGE:
DATE: / /

4. Scikit-learn
↳ Used for Machine learning coding in python.
↳ It was initially developed as a Google Summer of code project by David Cournapeau & released in 2007
↳ It is built on top of other python libraries so it provide full interoperationality with those libraries

5. ~~Then~~ TensorFlow
↳ It is free end-to-end open-source platform that has wide variety of tools, libraries and resources for AI.
↳ Developed by Google Brain team & released on November 9, 2015
↳ Using this we can easily build and train machine learning models with high level API's such as keras using TensorFlow.
↳ It allows to deploy machine Learning modles on cloud, browser or on your own device.

6. Keras
↳ Free, open source, neural network library written in python.
↳ Created by Francois Chollet, a Google engineer
↳ Released on 27 march 2015.
↳ Created to be user-friendly, extensible and modular while being supportive of experimentation in deep neural networks.

Rainbow
PAGE:
DATE: / . /

⊙ Python Libraries for Data Visulization.

1> Matplotlib
↳ Data visulization library and 2-D plotting library.
↳ Released in 2003
↳ used to create plots, bar charts, pie charts, histograms, scatter plots, error charts, etc.

2> Seaborn
↳ based on Matplotlib and closely integrated with the numpy and pandas data structures.
↳ It is high-level interface for creating beautiful and informative statistical graphics that are integral to exploring and understanding data.

3> Plotly
↳ Built on top of Plotly Javascript library
↳ used to create web-based data visulizations that can be displayed on Jupyter notebooks or web appln using Dash or HTML files.
↳ Provides 40 unique type of charts.

4> GGplot
↳ based on implementation of ggplot2 created of R lang.
↳ used to add different types of data visulization components or layers in a single visulization.

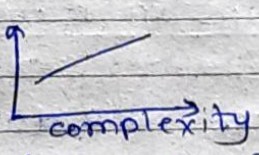
Rainbow
PAGE:
DATE: / /

* Data preprocessing

It is a technique that is used to convert the raw data into a clean data set.

* It includes:
1) Removing Duplicates
2) Transformation of Data using function or mapping
3) Replacing the values
4) Handling Missing Data.
5) None: Pythonic missing data.
6) NaN: Missing numerical data.

◎ Analytics Types



↳ Four types
↳ Interrelated and each of these offers a diff. insight

1) Descriptive Analytics (WHAT)
What is going on or what has happened.

2) Diagnostic Analytics (WHY)
↳ Focus on past performance to determine what happened and why.
↳ Helps in understanding the root of problem.

3) Predictive Analytics. (IF)
↳ Focuses on predicting the possible outcome using stastistical models and ML tech.
↳ History helps in prediction. / ↳ Forcasts.

4) Prescriptive Analytics:
↳ Recommands an action based on the forecast
↳ what should I do.

IMP

⊙ Apriori Algorithm.

⊙ ~~Apriori~~ Association Rule Mining algorithm. (ARM)

↳ Determines how strongly two objects are connected
↳ 'ARM' also called as Market Basket Analysis (MBA) and Affinity Analysis.
↳ Set of items in a transition is called Market basket.
↳ Mostly used in Retail as recommendation systems on ecommerse sites, etc..
↳ ~~If A and B~~
↳ If 'A' then 'B' {A ⇒ B} (↑ consequent) After A, B is bought
(↳ Product) (↳ Antecedent)
↳ In this algo BfS and Hash tree is used to calculate itemsets effectively.

* Support (S): Percentage (%) / number of times 'A' and 'B' both are accured in transitions.

('A' ⇒ 'B') = P(A∩B) { measures the frequency of association.

Rainbow
PAGE:
DATE: / /

('A' ⇒ 'B')

* Confidence (C): In a transition set 'T' if 'C' is the % of times 'B' is present in all the transitions containing 'A'

$$C = P(B/A)$$

$$C = \frac{P(A \cap B)}{P(A)}$$ } strength of association

conditional probability.

support of them {
- $P(A \cap B)$ → Number of times both occured
- $P(A)$ → Number of times A occured

* Parameters:
i> Finding all ~~associations~~ items that appears frequently in transitions } Min support count

ii> Finding strong associations among frequent items } confidence

* Problems in ARM
i> It is not easy to find level of ~~occu~~ frequency of appearance determination.
ii> It is not easy to find strong associations among frequent items.

* Functions of ARM
i> Finding ~~most~~ set of items that has significant impact on business.
ii> Collecting information from numerous tranctions.
iii> Generating rules from counts in tranctions.

* Strength of ARM
  i) Easy interpretation.
  ii) Easy to start
  iii) Flexible data formats
  iv) Simplicity.

* Weakness
  i) Exponential Growth in computations
     As transitions increases the number of combinations increases.
  ii) Lumps generation.
  iii) Problems in rule selection. confusion while selecting rule.
  iv) Not applicable for rare items. More applicable for frequent items.

* Apriori Algorithm:
↳ Idea is to generate candidate itemsets of a given size and then scan data set to check if their counts are really large.
↳ This process is iterative.

i) All singletone itemsets are candidates in the first pass. Any items will less than specified support value is eliminated.
ii) Two member item sets.
iii) Three member item sets,
iv) Frequent itemsets contains set of frequent itemsets.

v> Generate Association Rules which have confidence values greater than or equal to specified minimum confidence.

Ex.

Tid	items
1	2,3
2	1,3,5
3	1,4,2
4	2,3

min support 2

items	support
1	2
2	3
3	2
4	1
5	1

→ eliminated bec they are below support value

itemset	support
{1,2}	1
{2,3}	2
{1,3}	1

eliminated

Final set which we got {2,3}

Q. For the following Given Transitions Data-set, Generate Rules using Apriori Algo. consider the values as support : 50%, confidence : 75%.

Transition ID	Item
1	Bread, cheese, egg, Juice
2	Bread, cheese, juice
3	Bread, Milk, yogurt
4	Bread, Juice, Milk,
5	Cheese, Juice, Milk.

⇒ Freguent item set

$\delta = \frac{\text{Bread}}{n} = \frac{4}{5}$

items		Freguency		Support
Bread	→	4	→	4/5 = 80%.
cheese	→	3	→	3/5 = 60%.
⇒ egg	→	1	→	1/5 = 20%.
Juice	→	4	→	4/5 = 80%.
Milk	→	3	→	3/5 = 60%.
⇒ yogurt	→	1	→	1/5 = 20%.

egg & yogurt will get eliminated bec their support is below than given support.

* Make 2 item candidate set and write their freq.

item pairs	Freguency	Support
{Bread, cheese}	2	2/5 ⇒ 40%.
{Bread, Juice}	3	3/5 ⇒ 60%.
{Bread, Milk}	2	2/5 ⇒ 40%.
{cheese, juice}	3	3/5 ⇒ 60%.
{cheese, milk}	1	1/5 ⇒ 20%.

{juice, milk}	2	2/5 ⇒ 40%

Itemsets below support value 50%. get eleminated.

* For rules → {Bread, juice}
  → {cheese, juice}

(Bread, juice)
- (Bread ⇒ juice)
- (Juice ⇒ Bread)

(cheese, juice)
- (cheese ⇒ Juice)
- (Juice ⇒ cheese)

$$\text{confidence}(A \to B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

1) Bread ⇒ Juice $= \frac{S(B \cup J)}{S(B)} = \frac{3/5}{4/5} = \frac{3}{4} = 75\%$

2) Juice ⇒ Bread $= \frac{S(J \cup B)}{S(J)} = \frac{3/5}{4/5} = 75\%$

3) cheese ⇒ Juice $= \frac{S(C \cup J)}{S(C)} = \frac{3/5}{3/5} = \frac{3}{3} = 100\%$

4) Juice ⇒ cheese $= \frac{S(J \cup C)}{S(J)} = \frac{3/5}{4/5} = 75\%$

Following rules are generated:

(Bread ⇒ Juice)

(Juice ⇒ Bread)

(Cheese ⇒ Juice)

(Juice ⇒ cheese)

Q. Generate rules using Apriori algo. Values support 22% and confidence = 70%.

Transaction ID	Items Purchased
1	I_1, I_2, I_5
2	I_2, I_4
3	I_2, I_3
4	I_1, I_2, I_4
5	I_1, I_3
6	I_2, I_3
7	I_1, I_3
8	I_1, I_2, I_3, I_5
9	I_1, I_2, I_3

⇒ Frequent itemsets

Item	frequency	support
I_1	6	6/9 ⇒ 66%
I_2	7	7/9 ⇒ 77%
I_3	6	6/9 ⇒ 66%
I_4	2	2/9 ⇒ 22%
I_5	2	2/9 ⇒ 22%

None of item is eliminated because all have specified support value.

2) Let make group of Two item

Itemsets	Frequency	Support
(I_1, I_2)	4	4/9 ⇒ 44%.
(I_1, I_3)	4	4/9 ⇒ 44%.
(I_1, I_4)	1	1/9 ⇒ 11%.
(I_1, I_5)	2	2/9 ⇒ 22%.
(I_2, I_3)	4	4/9 ⇒ 44%.
(I_2, I_4)	2	2/9 ⇒ 22%.
(I_2, I_5)	2	2/9 ⇒ 22%.
(I_3, I_4)	0	0
(I_3, I_5)	1	1/9 ⇒ 11%.
(I_4, I_5)	0	0

3) Let make group of Three item

Itemsets	Frequency	support
(I_1, I_2, I_3)	2	2/9 → 22%. ✓
(I_1, I_2, I_4)	1	1/9 ⇒ 11%.
(I_1, I_2, I_5)	2	2/9 ⇒ 22%. ✓
(I_2, I_3, I_4)	0	0
(I_2, I_3, I_5)	1	1/9 ⇒ 11%.
(I_1, I_3, I_4)	0	0
(I_1, I_4, I_5)	0	0

(I_1, I_3, I_5)	1	$1/9 \Rightarrow 11\%$.
(I_2, I_4, I_5)	0	0
(I_3, I_4, I_5)	0	0

~~(I_3, I_4, I_2)~~

~~$(I_5$~~

Two itemset have support value greater than given value. Those set are (I_1, I_2, I_3) and (I_1, I_2, I_5)

*For rules $\rightarrow (I_1, I_2, I_3)$
$\rightarrow (I_1, I_2, I_5)$

$$\text{confidence} = \frac{\text{supp. } (A \cup B)}{A}$$

	confidence	
$(I_1, I_2) \rightarrow I_3$	2/4	$\Rightarrow 50\%$.
$(I_1, I_3) \rightarrow I_2$	2/4	$\Rightarrow 50\%$.
$(I_2, I_3) \rightarrow I_1$	2/4	$\Rightarrow 50\%$.
$(I_1, I_2) \rightarrow I_5$	2/4	$\Rightarrow 50\%$.
$(I_2, I_5) \rightarrow I_1$	2/2	$\Rightarrow 100\%$. ✓
$(I_1, I_5) \rightarrow I_2$	2/2	$\Rightarrow 100\%$. ✓
$I_1 \rightarrow (I_2, I_3)$	2/6	$\Rightarrow 33\%$.
$I_2 \rightarrow (I_1, I_3)$	2/7	$\Rightarrow 28\%$.
$I_3 \rightarrow (I_1, I_2)$	2/6	$\Rightarrow 33\%$.
$I_5 \rightarrow (I_1, I_2)$	2/2	$\Rightarrow 100\%$. ✓

Following Rules are generated
$(I_2, I_5) \rightarrow I_1$,
$(I_1, I_5) \rightarrow I_2$,
$I_5 \rightarrow (I_1, I_2)$

Rainbow

PAGE:
DATE: / /

* FP Growth

(Frequent Pattern Growth)

↳ classicial algo. in data mining.

↳ Divide and ~~co~~ conquer

FP-Growth algo

↳ Compresses data sets to a FP-tree.

↳ Scans the database twice.

↳ Gently improves mining efficiency.

Ex.

TID	Itemsets
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

min support [3]

⇒

Items	Support
M	3 (No of times appeared)
O	3
* N	2
K	5
E	4
Y	3
* D	1
* A	1
* U	1
* C	2
* I	1

L

K	5
E	4
M	3
O	3
Y	3

↳ frequent Pattern.

Rainbow

PAGE:
DATE: / /

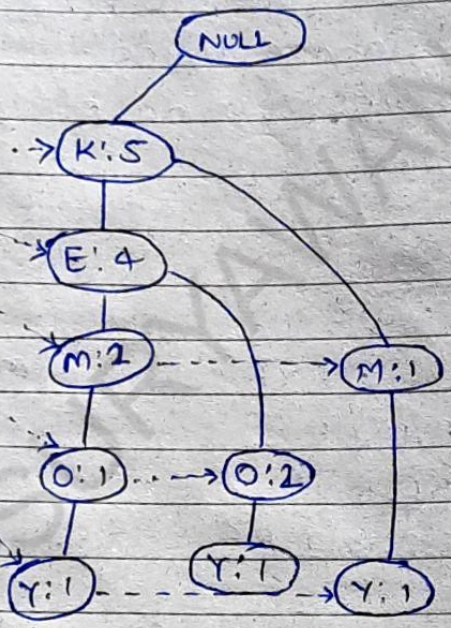
original writing itemset into ordered according to FP. write using ~~as~~ assending order.

write according to frequent patt.

Ordered set
K, E, M, O, Y
K, E, O, Y
K, E, M
K, M, Y
K, E, O

Build FP-tree using ordered set.

	S.C.	N.L
K	5	
E	4	
M	3	
O	3	
Y	3	



write according to bottom to up from FP tree

Items	Conditional patt. base (path)	Conditional FP tree (common items)
Y	{(KEMO:1), (KEO:1), (KM:1)}	{K:3}
O	{(KEM:1), (KE:2)}	{KE:3}
M	{(KE:2), (K:1)}	{K:3}
E	{(K:4)}	{K:4}
K	-	

Rainbow

PAGE:
DATE: / /

combination of 1-item FP tree.

	Frequent patt. Generated
Y	<K,Y:3>
O	<K,O:3> <E,O:3> <O,E,K:3>
M	<M,K:3>
E	<E,K:4>
K	

Q.

T.R Id	Items
1	E,A,D,B
2	D,A,C,E,B
3	C,A,B,E
4	B,A,D
5	D
6	D,B
7	A,D,E
8	B,C

Min supp. : 30%.

→

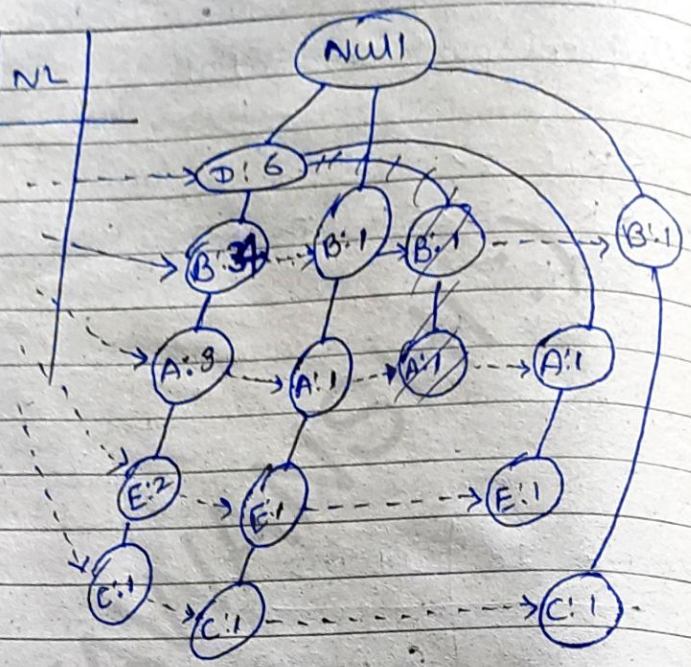
Item	Support.
A	5/8 = 62.5%.
B	6/8 = 75%.
C	3/8 = 37.5%.
D	6/8 = 75%.
E	4/8 = 50%.

~~ordered set~~

Frequent Pattern

D	75%.
B	75%.
A	62.5%.
E	50%.
C	37.5%.

Ordered set		SC.	NL
D,B,A,E	D	6	
D,B,A,E,C	B	6	
B,A,E,C	A	5	
D,B,A	E	4	
D	C	3	
D,B			
D,A,E			
B,C			



Big Data Analytics & Model Evaluation. Rainbow

PAGE:
DATE: / /

* **Clustering:**
  - ↳ unsupervised learning technique.
  - ↳ Grouping objects of similar type / objects sharing similar characteristics.
  - ↳ used when we want to explore data.
  - ↳ 'clustering algo.' used for natural grouping
    - → Hierarchial — cluster within cluster. e.g. sports news → Cricket, Baseball
    - → Partitional — Fixed number of clusters. e.g. K-means.

* **K-means** → averaging data / Finding centroid
  - ↳ Number of clusters.
  - ↳ Method of vector quantization

2.36 → 2
2.78 → 3

The diff after rounding off is called quantization

- → Partition of 'N' observations in 'k' clusters.
- ↳ Each observation belongs to cluster with nearest mean, serving as prototype of cluster.

? Steps
- ↳ Number of clusters required (like 1, 2, etc.)
- ↳ Assigning the cluster to each element depending on min distance.
- ↳ Each time new element added to the cluster, the centroid position get recalculated. This process is performed until all the elements are grouped in cluster.

Rainbow
PAGE:
DATE: / /

* Centroid → Point which represents mean of the parameter values of all the pts in cluster.

* Objective
  - ↳ Group similar data points together and discover underlying patterns.
  - ↳ Finds fixed number of clusters from dataset. which is 'K' cluster.
  - ↳ K-means algo. finds the 'K' number of clusters/centroids & then allocate the data points to the nearest cluster. while keeping centroid value as low as possible.

* Working:
  - ↳ starts with grouping randomly selected centroids, used are used as begining pt. of every cluster
  - ↳ Then we perform iterative operations for optimising the centroid.
  - ↳ It stops when
    1. Centroid have stabilised the value, no change in value as clustring is successful.
    2. Defined no of iterations have been achives.

* Use
  - ↳ Finding groups form the data which is not labled.
  - ↳ data cluster analysis

Rainbow

PAGE:
DATE: / /

Advantages:

1) It easily adpts to new examples.
2) It can generalise to clusters to diff. shapes, sizes such as elliptical clusters.

Dis adv:-

1) Difficult to predict K-value.
2) Does not work better with global cluster.
3) Diff. initial patterns can give diff. results.

* Methods of using ~~clus~~ K-means

> step 1 → choose the number of cluster (K) → like (2)
> step 2 → select ~~the~~ 'K' random data points as centroid.
> ~~step 3 → A~~

Brand	Price	ram
	45	8
	17	5
	30	6
	15	4
	12	4

step 1 →

	Initial centroid	
cluster	price	ram
K1	45	8
K2	17	5

> step 3 → Assign points to the closest cluster.
↳ calculate Euclidean distn betn each of these 2 cluster centroids and each of observations.

Consider this from dataset table

Data Ent.	Cluster	price	Ram
1	K_1	45	8
2	K_2	17	5

Rainbow
PAGE:
DATE: / /

Euclidean dist^n $\rightarrow \sqrt{(X_H - H_1)^2 + (X_W - W_1)^2}$

specified type here (price) (Ram)

$X_H \rightarrow$ centroid value
$H_1 \rightarrow$ observation value
$X_W \rightarrow$ centroid value
$W_1 \rightarrow$ observation value

calculating for random value for K_1 cluster

(2) as per table $\Rightarrow \sqrt{(45-17)^2 + (8-5)^2}$

$\Rightarrow$ ~~28+3~~ $\Rightarrow$ ~~31~~ 28.16

(1) as per table value for K_2 cluster

$\Rightarrow \sqrt{(17-45)^2 + (5-8)^2}$

$= -28 - 3 \Rightarrow$ ~~-31~~ 28.16

(1) as per table

$= \sqrt{(45-45)^2 + (8-8)^2}$

$= 0$

so | 45 | 8 | data set value will go in (1) cluster bec it have lowest value in that cluster.

> step 4 → Recompute centroid or as per book method

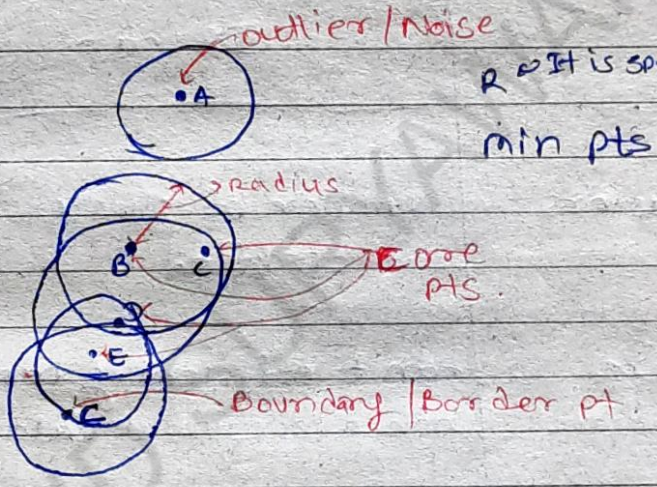
1) calculate all Euclidean values and then average out and get new centroid value according that.

* DBScan Clustering

(Density based Clustering of Appl^n with Noise)

↳ The neighbourhood of each point in a cluster which is within a given Radius must have a minimum no. of pts.

↳ effecient for detecting outliers / noise.



Algo. DBScan :-

Data point in data set will be of following any type.

* **Core Point :** If there are min number of points within the specified radius.
  Like. Three min pts so here B, C, D, E are core pts.

* **Border point →**
  ↳ contains less than min pts.
  ↳ should be in radious of neighbour core pt.
  Here C

Outlier pt :> Not close enough to reach core point.
These pts get eleminated

> cluster of core and border points get formed.

* Diff. betn K-means and DBScan

K-means	DBScan
1> Number of clusters are specified	Number of clusters are not specified
2> cluster is more efficient.	can't efficiently handle high dimentional data sets.
3> Does not work well with outliers	3> works well with outliers and noisy data.
4> Require one parameter number of cluster(K)	4> Require two parameter i> Radius(R) and ii> Minimum points (M)

Rainbow

PAGE:
DATE: / /

Read topics till Bag-n-gram from book.

* Term Frequency - Inverse Document Frequency (TFIDF)

↳ It is used for information retrival and text analysis.
↳ This is similar to the how google search engine works.

when we search something that is called as query Google search that query in the aralible html docs and shows the result.

TFIDF = Term Freq. (TF) X Inverse Doc. Freq. (IDF)

TF → Frequency of word in given doc.
IDF → the measure of the importance of word.

(Term Freq.)
$$TF = \frac{\text{count of term in document}}{\text{Total number of terms in that doc}}$$

How many time word come in that doc / Total words in that doc

$$IDF = \log\left(\frac{\text{Number of total Doc}}{\text{Number of documents containing that word}}\right)$$

Q. How 'fox' word is relerent in following docs.

Doc 1 → A quick brown fox jumps over the lazy dog. what a fox!

Doc 2 → A quick brown fox jumps over the lazy fox. what a fox! what a man.

⇒ Doc. D_1 have 12 words
Doc. D_2 have 15 words

Calculate freq of 'fox' in both doc.

Term Freq. (TF_1) (D_1) $= \frac{2}{12} = \frac{1}{6}$

Term freq. (TF_2) (D_2) $= \frac{3}{15} = \frac{1}{5}$

* IDF (Inverse doc freq) (D_1) $= \log\left(\frac{\text{Total no of Doc}}{\text{Times word occ in Doc } D_1}\right) = \log\left(\frac{2}{2}\right) = 0$

IDF $= \log\left(\frac{\text{Total no. of Doc}}{\text{Times word occ. in Doc } D_2}\right) = \log\left(\frac{2}{3}\right) =$

IDF (Inverse Doc Freq.) $= \log\left(\frac{\text{Total no of Doc}}{\text{word in number of doc.}}\right) = \log\left(\frac{2}{2}\right) = 0$

There are two doc then fox word is in 1 doc or 2 doc that is need to write

$TFIDF = TF_1 \times IDF = \frac{1}{6} \times 0 = 0$

$TFIDF = TF_2 \times IDF = \frac{1}{5} \times 0 = 0$

As the TFIDF score is same for both doc the 'Fox' word have same/ equally relevent for both doc.

Rainbow

PAGE:
DATE: / /

5.8. Social Network Analysis (SNA)

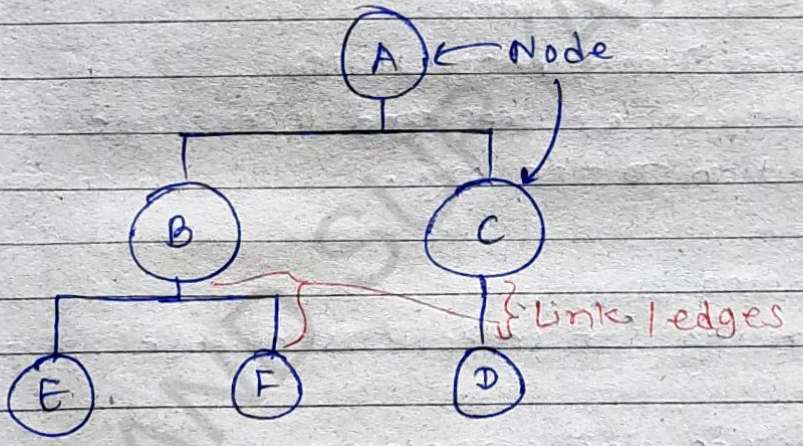
↳ uses graph theory model.

↳ It is process of investigating social structures through use of social networks / graph theory.

↳ Each graph has nodes and edges

nodes → object in graph

edges → Link which shows the relⁿ betⁿ the obj in graph.



> Degree of Node : No. of edges it have
A degree is 2 , D degree is 1.

> Path length : Distⁿ betⁿ two nodes
i> betⁿ A and D → A → C → D ∴ Path len. is 2
ii> betⁿ F and D → F → B → A → C → D ∴ Path len is 4

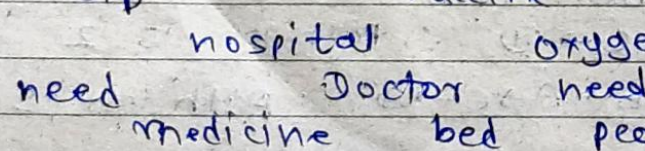
> Centrality → Importance of particular node in the graph / network
↳ key node that join several networks.
↳ Here A is can be treated as high centrality bec it joins two groups B & C

> Density → It refers to the proportion of direct connection to the total no. of direct conn. possible in network.

* Need of Social Network Analysis.
1) Identifying the Influencers.
2) Human Resource Management (HRM)
↳ identify critical resources and understand their contribution to the org.

3) contact tracing
↳ contact racing for infectious diseases.

4) Identify themes and connections.
↳ identify dominant themes & relations betⁿ keywords and identify sentiments.
e.g. Journal of medical Internet Search shows con betⁿ 10 words



Rainbow
PAGE:
DATE: / /

5. Fraud detection
↳ can detect fraud connections ~~which~~ which have criminal rec.

* Intro to Business Analysis.
Doing some change in organisational context (changing something/ guiding) by defining need & recommanding solutions that can produce some value or growth to stake holders.

↳ Business Analysis helps organisations to do business better.

⊛ Skills need for Business Aanalyst role.
↳ Analytical skills
↳ Communication skills
↳ knowledge/understarding of that business
↳ Experience in handling projects
↳ Descision making.
↳ Influence
↳ connections/management
↳ Presentation, ↳ Problem solving etc.

⊛ Dealing with key Stakeholders.
↳ major source of requirements, constraints and assumptions for the Business Analyst.

Some important stake holders
↳ customer
↳ Domain subject matter expert.

↳ End user
↳ Projece manager
↳ Tester
↳ Sponsor
↳ Supplier
↳ Regulator.

5.10. Model Evaluation and selection

A) Cross validation

techq. to evaluate performance of machine learning algo. By using the avalible input data training the avalible ML modles and evaluating them the complementary subset of data.

⊙ Types

i) Non-Exhaustic:
↳ split orignal dataset and leave out subset for validation.
↳ Do not attempt to split data set in all possible ways.

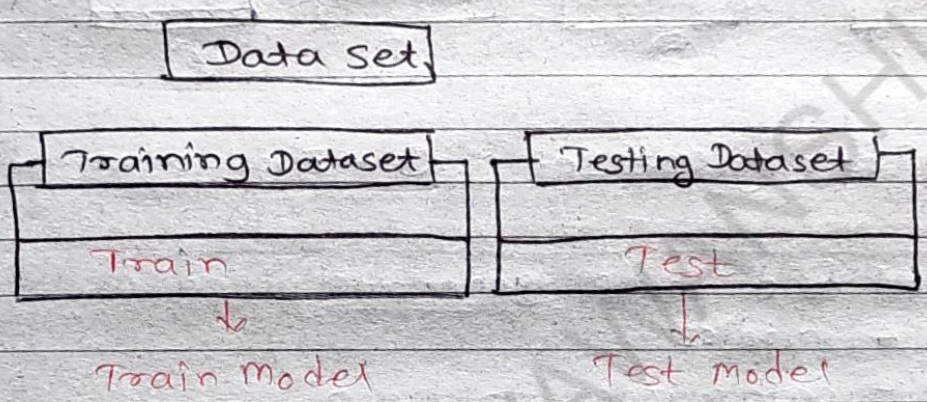
ii) Exhaustic
↳ You split data set into all possible ways and into training and a validation set and carry out multiple iterations of cross-val.

Rainbow
PAGE:
DATE: / /

(B) Holdout Method.

↳ Is a non-exhaustic approach.

↳ In this approach data set is divided in two parts.



↳ size of ~~testing~~ training data is more than twice of ~~training~~ testing data.

↳ Ratioms 70:30 and 80:20 are common here.

↳ Data is shuffled before spliting.

↳ Just one iteration is performed.

↳ It may achieve highly missleading results.

↳ "often known as simplest kind of cross validation but not great cross validation tech.

(C) K-Fold Cross validation.

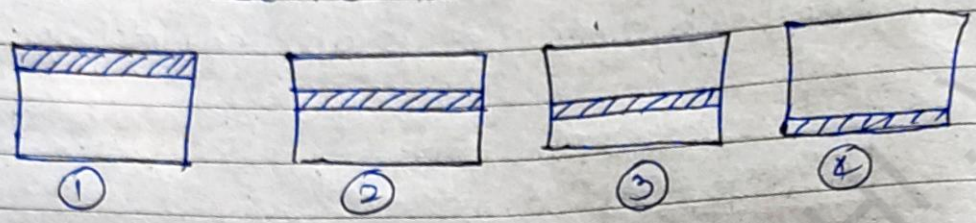
↳ Non-exhaustive app.

↳ Split the input data into diff 'K' subsets of data also called folds.

↳ You train machine learning model on all but one (K-1) subsets and then evaluate model on the basis of subset that is excluded.

used to evaluate

↳ Process is repeated several times, with diff subset reserved for evaluation each time.



O → Training dataset ▨ → Evaluation dataset.

↳ This 4 fold model generate
- ↳ 4 modles
- ↳ 4 datasources to train
- ↳ 4 datasources to evaluate one for each

① Leave-P-Out Cross-Validation (LpOCV)
- ↳ exhaustic approcach.
- ↳ 'p' numberof points are taken out from the total number of data points in the dataset (say n)
- ↳ while training model you use these (n-p) points to train the model.
- ↳ The process is get repeated for all the possible combinations of p.

Ex. Given dataset
Fruits = {apple, orange, guava, grapes, mango}
If you choose p=2
↳ For each Iteration 2 points are keep away

Rainbow
PAGE:
DATE: / /

↳ To ensure all the possible combinations of p, we split the dataset into iterations.

⊙ Iteration 1:
Training = {apple, orange, guava}
validation = {gropes, mango}

⊙ Iteration 2
Training = {guava, gropes, mango}
validation = {apple, mango}

many more can possible....

Total no. of combination can calculated by

$$C = \frac{n!}{p!\,(n-p)!}$$

⊛ LpOCV with p=2 → Leave pair out cross validation
⊛ LpOCV with p=1 → Leave one out cross validation

⑤ Sub-sampling
↳ technique using which you can use multiple subsets of a large dataset for taining
↳ To reduce time and increasing performance
↳ seep up model training ~~and~~ while maintaining the model performance.
↳ It allows to specify a percentage of training data to be used for training using configurable hyperparameter.

Rainbow

PAGE:
DATE: / /

↗ It is placed / put manually.

hyperparameter: It is a parameter whose value is used to control the learning process.

↳ Hyperparameter is used for control various aspects of training a machine learning model.

* Hyperparameter tunning.
  ↳ It is the process of determining the right combination of hyperparameters that allows the model to maximise model performance.

* Hyperparameter tuning Technique.

(a) Grid Search.
  ↳ In grid search approach, machine learning model is evaluated for a range of hyperparameter values.
  ↳ It is called as grid search bec it search for best set of hyperparameters from a grid of hyperparameter values.

(b) Random Search.
  ↳ Insted of searching over the entire grid random search only evaluates a random sample of points on the grid.
  ↳ It is alot cheaper than grid search.

Rainbow
PAGE:
DATE: / /

(c) Bayesian Optimisation.
↳ Helps to find the minimal point in the min. no of steps.
↳ uses an aquisition function that directs sampling to areas where an improvement over the current best observation is likely.

(d) Tree-Structured Parzen estimators (TPE)
↳ It is similar to Bayesian optimisation.

Data Visulization & Hadoop

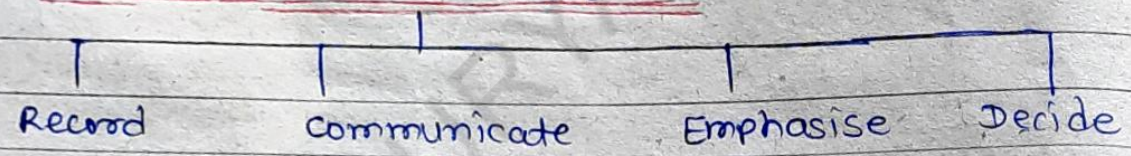
Rainbow
PAGE:
DATE 13/06/2022

* Data Visulisation

↳ It is a graphical or pictorial representation of data that makes it easy to communicate the info. to humans. (understand)

↳ uses various forms of representations to match the data and the relationship amongst its data attributes so as to communicate the desired info effectively.

* Goal of data visulisation.



1) **Record**:
   ↳ Helps to record info in various forms such as tables, logs, emails, audio, video or any other form of info sharing.
   ↳ User may have to dig through several info to find / get desired data.

* 2) Communicate

↳ communicate the info in most effective way.
↳ we can use charts, maps and graphs effectively to visulise different forms of data as well as relationship amongst its data attributes.

Rainbow
PAGE:
DATE: / /

3> Emphasise :

↳ emphasise or highlight a portion of data, find patterns, show trends or depict relⁿ betⁿ various data attributes.

↳ Like showing view of rainfall in various states
- Excess 20% or more
- Normal 19% to -19%.

After putting or marking regions using this it will give us proper understanding.

4> Decide

↳ Data visulisation makes it easy to quickly make decisions and take actions.
↳ we do not have to go through the length reports & complex data to understand what need to be done.
↳ But representing data visually help us to make / take decision quickly.

* Challenges with Big Data Visulisation.

major challenges
- ↳ Nature of big data
- ↳ Lack of experts
- ↳ Cognitive challenges.

Rainbow

PAGE:
DATE: / /

1] Nature of Big Data:

↳ The massive volm of data requires special software and hardware to visulise it.

↳ The variety of data makes it hard to cenceptulise the right form of visulisation to show relationship betn data attributes.

↳ As velocity of data is high! you have to contineously update your visulisation to keep it accurate.

↳ The veracity and value characteristics require that your visulisation meets the data quality and usefulness as well.

↳ If the visulisation is formed with poor quality of data then it may not fulfill the desired objectives of visulisation.

2) Shortage of Expert:

↳ It is an emerging field (data analytics) so there are less amount of experts around world.

↳ Experts are required who can

a) Understand wide variety of data.

b) Model data correctly.

c) Build and manage hardware and software tools for big data visulisation.

d) Design appropriate visual interfaces.

3) Cognitive challenges.

↳ Irrespective of what we have got ultimately we have to understand that from those visulisations like pie chart, graps, curves etc.

↳ Also have to deal with other visulisation errors and omissions, illusions, phases, perception and short attention span challenges.

* Techniques for Visual Data Representation.

⊙ Data visulisation techniques are choosen on bases of
> The audience who have to understand that data.
> Type of data.
> Range of data.
> Purpose
> context setting or the oretall big picture you want to draw.

⊙ Conventional data visudisation tools such as tables, bar chart, pie chart, etc. that were previously used to visualise the info.

⊙ Types of data visulisation

At highlevel various data visulisations could be grouped as.

Types of data

Rainbow

PAGE:
DATE: / /

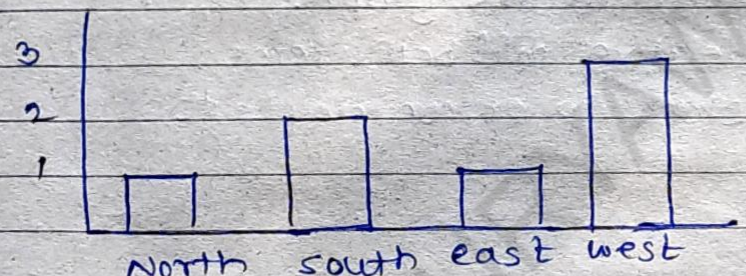
Types of data visualisation

- Comparative Plots
  - i> column and Big charts
  - ii> Line chart
  - iii> Area chart
  - iv> Bubble chart
  - v> Pie chart
- Statistical plots
  - Histograms
  - Scatter plots
  - Box plot
  - Radar plot
  - Tree Map
  - Waterfall chart.
- Topology Plots
  - Linear Topology
  - Graph Topology
  - Tree Topology
- Spatial Plots.
  - choropleth. Map
  - Point Map
  - Raster surface
  - Heat Map
  - World Cloud.

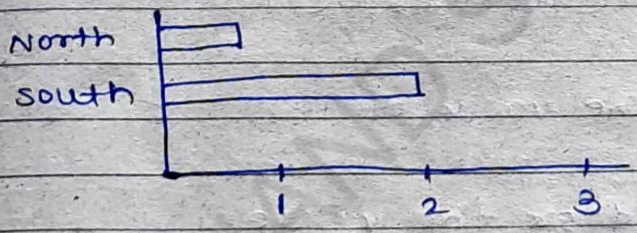
Rainbow
PAGE:
DATE: / /

A) Comparative Plots
Are used for comparing the datapoints.

1) Column and Bar Charts.
↳ Used to compair two or more values in same category.
↳ There are several variations like column chart, bar chart, stack chart and their 2D and 3D plots.



column chart



Bar chart.

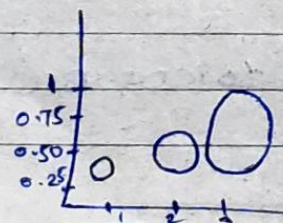
↳ These are easy to read and understand and indivisual values ~~ehr~~ can be changed without affecting others.
↳ But this do not work well if there are several categories.

2) Line chart.
↳ used to show time-series data and to compair categories.

↳ used to understand trends and patterns in your data and also make future projections.
↳ minor variations in line chart.
e.g. sensex chart.
↳ For showing variation in few categories, Line chart works well but for more categories it is not good.

3) Area ~~Graph~~ chart.
↳ It is very similar to line graph.
↳ Heighlights the relative diff. betn items.
↳ Different item stack up or contribute to the whole.

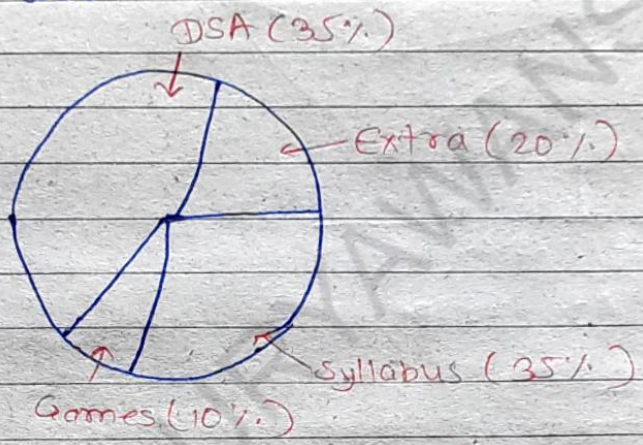
4) Bubble Chart
↳ It can show comparison as well as distribution
↳ Size of bubble shows the value of the datapoint.
↳ Bigger the bubble higher the value it represents.
↳ Bubbles are plotted by three different values.
- one value for position along x-axis
- one value for position along y-axis
- one value for size of bubble.



Rainbow
PAGE:
DATE: / /

5) Pie chart

↳ One static member that is divided into several categories that contains its individual portion.
↳ Each portion is some % amount of whole.
↳ When we sum up all the portions value then that comes 100%.



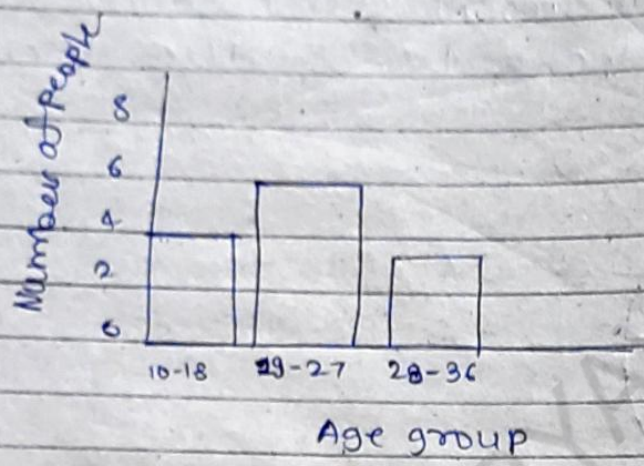
↳ It is useful when you want to show breakdown of data into several categories and their relⁿ ship to the whole.
↳ Not suitable for more categories bec it will become indistinguishable.

B) Statistical Plots.
↳ useful to show results of statistical analyses.

1) Histograms

↳ represents the frequency of the dataset in a dataset as well as its distribution.

↳ Looks similar to the column chart but it is diff. As it plots the frequency for each distribution rather than the actual value of any datapoint itself.

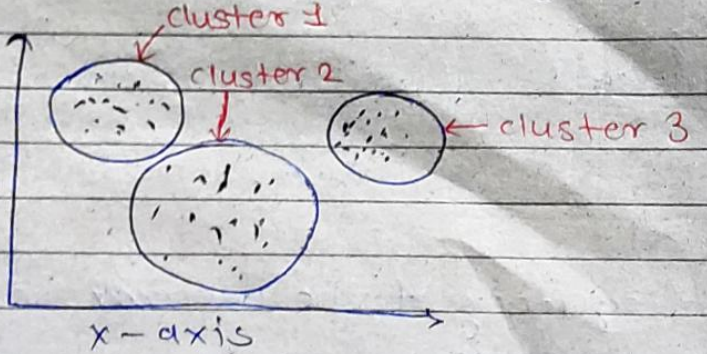


2) Scatter plot.

↳ used to find trends, clustering and outliers from a given dataset.

↳ Datapoints are plotted according to their coordinate values to reveal patterns.

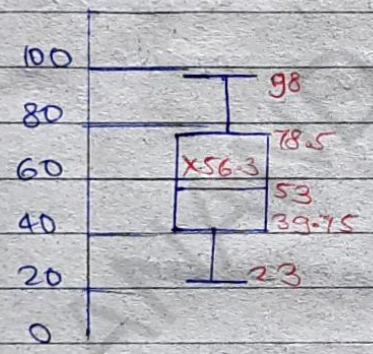
↳ useful when looking for outliers or for understanding the distribution of your data.



3. Box Plot (Box and whisker diagram)

↳ shows how datapoints in dataset are distributed
↳ It shows several values at once (statistical)
↳ some statistical values that are shown are
- Minimus
- 1st Quartile
- Mean
- Median
- 2nd Quartile
- 3rd Quartile
- Maximum

Ex. Marks of Maths of class A



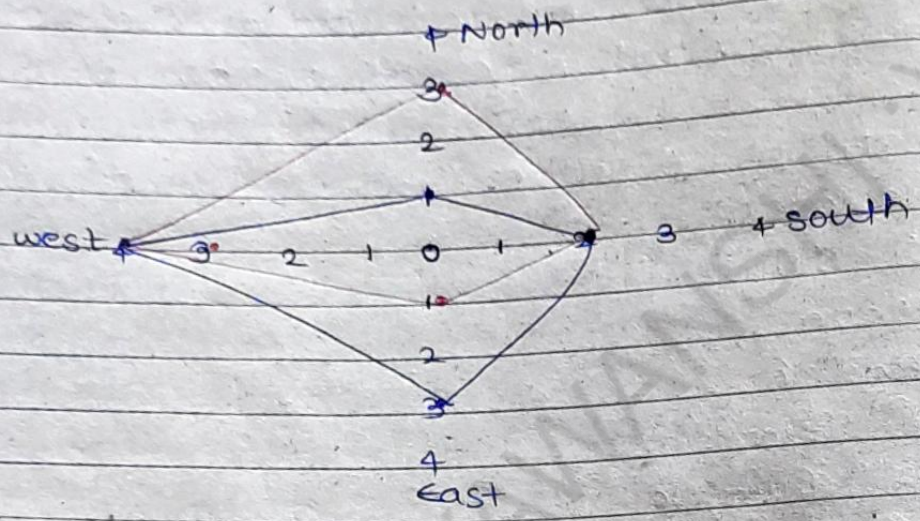
Minimum → 23
1st Quartile → 39.75
Mean → 56.3
Median → 53
2nd Quartile → 53
3rd Quartile → 78.5
maximum → 98

* 4) Radar chart.

↳ Used to understand relative diff betⁿ datapts.
↳ Helps to highlight the relative diff betⁿ the datapoints
↳ Helps to focus on major differences.

Sales Region	A	B
North	1	3
South	2	2
East	3	1
West	4	4

Gr. A ⇒
Gr. B ⇒



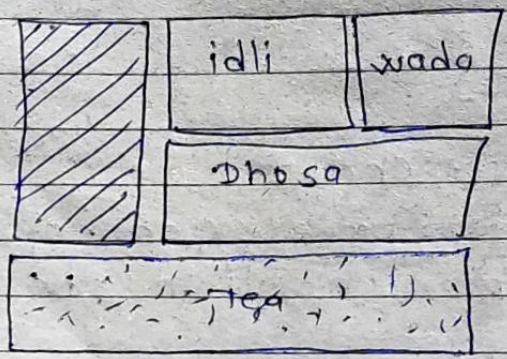
5. Tree Map.

↳ useful to show aggregate parameters of similar categories and then use area to show the relative size of each category compare to the whole.

↳ Provides hierarchical view of data.

e.g. Sales of Nasta shop.

Tea ▨ other ▨ South indian □

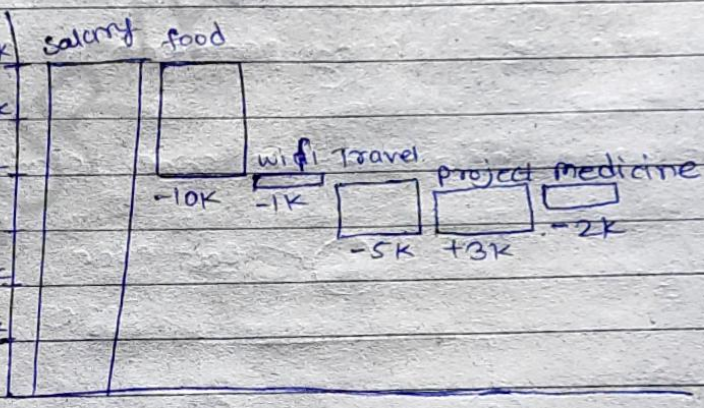


Rainbow
PAGE:
DATE: / /

6. Waterfall chart
↳ shows how a particular value is affected by intermediate values.
↳ It can be -ve or +ve
↳ Each datapoint considers automatically all the previous datapoints values / plots.
↳ It is easy to understand how the particular value is changing with each datapoint on chart.

Eg. Monthy Expense

Objective	Trans.
salary	30,000
Food	-10,000
wifi	-1,000
Travel	-5,000
project	+3,000
Medical	-2,000



↳ Topology plots.

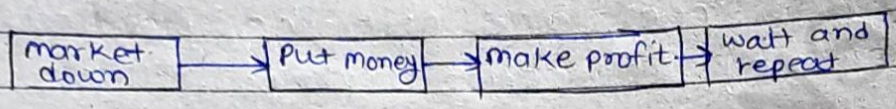
↳ Uses geometrical structures to show relationship and connectedness betⁿ datapoints in dataset.
↳ Used to show relationships, hierarchies, connectedness etc. for datapoints which do not have numerical values associated with them.

Rainbow

PAGE:
DATE: / /

1> Linear Topology

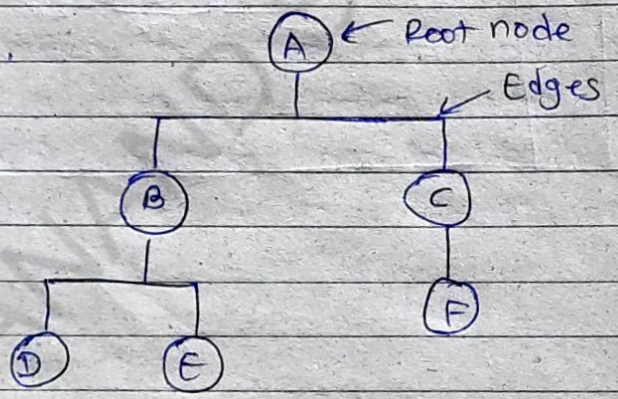
↳ shows one-to-one rel^n bet^n entities.
↳ Shows process / steps to achive a goal or perticular outcome.
↳ E.g. Share Market



2> Graph Topology

↳ Shows multiple rel^n bet^n datapoints
↳ Highly used in social media connection analysis, finding pattern in com^n and info flow
↳ Shows how entities are related.

e.g.



↳ Becomes dense as nodes and edges increases

3> Tree Topology

↳ represents hierarchial classification.

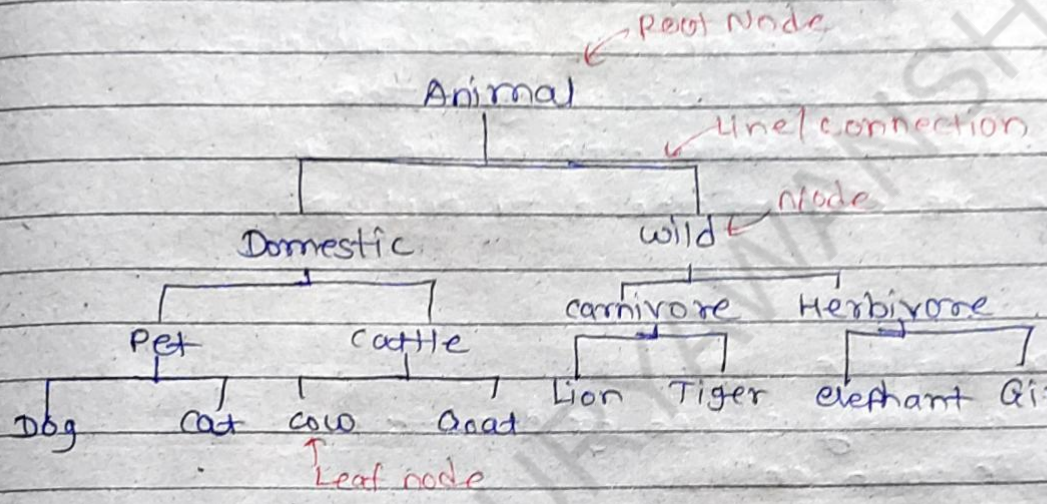
Rainbow

PAGE:
DATE: / /

↳ Root node begins the tree followed by several branches and leaves.

↳ Node → recievers
  ↳ distributors of connections

Line → con^n bet^n nodes.



2] Spatial plots

↳ use logical space for visulisating the data

↳ Logical space could represent map, location, shape or size of data attribute.

⇒ Choropleth Map

↳ It uses diff. shading, coloring or the placing of symbols within predefined areas to indicate the average values of a particular quantity in those areas.

↳ The datapts are plotted according to area boundary.

↳ e.g. Litracy rate in India as per Census 2011 shown on India's map using shading of colour.

2> Point Map
↳ Point map uses a symbol (bubble, etc.) to highlight the datapoints and coverage.
↳ Points are used to highlight the distribution of data and as well as respective size of the distribution.

3> Raster Surface
↳ It is a evenly spaced grid containing a value in each cell.
↳ It could be used for satellite img map to a surface coverage map with values that have been interpolated.
↳ A surface can have infinite number of pts with various values.

4> Heat Map
↳ similar to choropleth map.
↳ no need of geological boundary or region.

5> Word cloud
↳ similar to heat map.

↳ It uses font size and colours to highlight the most prominent textual data.
↳ Used to highlight the keywords which are being mostly used.

⊙ Data Visualisation Taxonomy.
(way to classify the different kind of plots.)

Data visualisation Taxonomy
- 1. 1D (Linear)
- 2. 2D (Planer)
- 3. 3D (Volumetric)
- 4. Temporal
- 5. nD (Multidimensional)
- 6. Tree (Hierarchial)
- 7. Network.

1. 1D (Linear) Visualisation

↳ linearly or sequentially visualised.
↳ These can be textual documents, program source code or just alphabetical lists of names which are all organised in sequential manner.
↳ simple listing organised in perticular way.
↳ Generally it is not graphically visualised.