Unit - 1

Facets of Data

- Structured Data
- Unstructured Data
- Natural language
- Machine Generated
- Graph - based
- Audio, video and Images
- Streaming.

→ **Structured Data:** (Predefined) Eg: Excel table [rows & columns Contains date, name address, etc.]

* Structured Data is data that depends on a data model and resides in a fixed field within a record.

* It is easy to store Structured data in tables with in databases or Excel files.

* Sql (or) Structured Query language is the preferred way to manage and query data that resides in DB.

→ **Unstructured Data:** (not predefined) includes Text & multimedia content.

* Unstructed data is data that is not easy to fit into a data model because the Content is Context Specification (or) Varying.

Eg: Email, Satellite images, Radar, Social media data, website content

→ **Natural Language:**

* Natural language is a special type of Unstructed data.

It is challenging or difficult to process because it requires knowledge of specific data science techniques and linguistics.

Eg:- Email, word doc.

Natural Language processing techniques are, Topic recognition, Summarization, text completion & sentiment Analysis.

(4) Machine generated Data:

Information that is automatically created by a computer, process, application or other machine without human intervention.

Eg: Web server logs, call detail records, n/w event logs.

Analysis of machine data relies on highly scalable tools due to its high volume & speed.

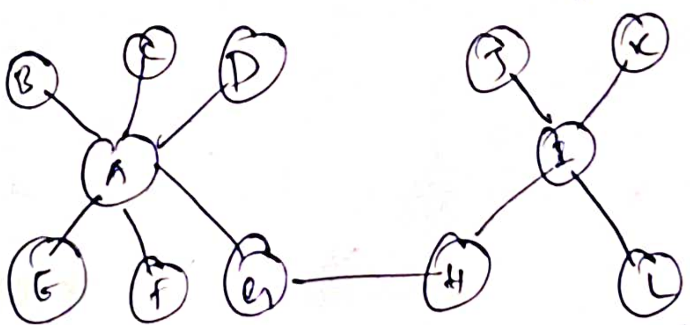
(5) Graph based (or) n/w based:

* Graph (or) N/w data is data that focuses on relationship b/w objects.

* Graph structures uses nodes, edges and properties to represent and store graphical data.

* Graph based data is a natural way to represent social n/w's.

Eg:



friends in Social n/w's. (Twitter, fb).

6) Audio, Image & video.

Multimedia data in the form of audio, video, images become an integral part of everyday life.

Object recognition - Challenging for computers,

Deep mind - developed an algorithm which is capable of learning how to play video games - able to interpret everything in the video Screen via deep learning.

7) Streaming data:

It takes almost any of the previous forms but it means the data flows into the system when an event happens instead of being loaded into a data store in a batch.

Eg: "what's trending" on Twitter, live sporting (or) music events.

The DATA SCIENCE PROCESS

It Consists of 6 steps. They are

1. Setting the Research goal.
2. Retrieving data
3. Data preparation
4. Data Exploration
5. Data Modeling
6. Presentation & automation.

1. Setting the Research goal: (acquiring data)

→ The purpose of this step is making sure all the stake holders understand the what, how, and why of the project.

2. Retrieving data: (Collection of data which acquired for project)

→ This step includes finding suitable data and getting access to the data from the data owner.

→ This results in data in its raw form.

3. Data preparation: This includes transformating the data from a raw form into usable form (data).

→ This involves detecting and correcting different kinds of error in the data, & Combine data from different data sources & transform it.

④ Data Exploration:

This step involves finding patterns, correlations and deviations based on visual and descriptive techniques to gain a deep understanding of the data.

⑤ Data Modeling

Select the variables to built the model & also a modeling technique.

Building models with the goal of making better predictions, classifying objects.

⑥ Presentation and automation:

Finally present the result to the business. Results can take many forms ranging from presentation to research reports.

Automate the execution of the process if needed.

Overview of the DS process :

Step-1 : Refining research goals and Creating a project Charter

* Understanding the what, why and how of the project & answering questions is the goal of 1'st phase.

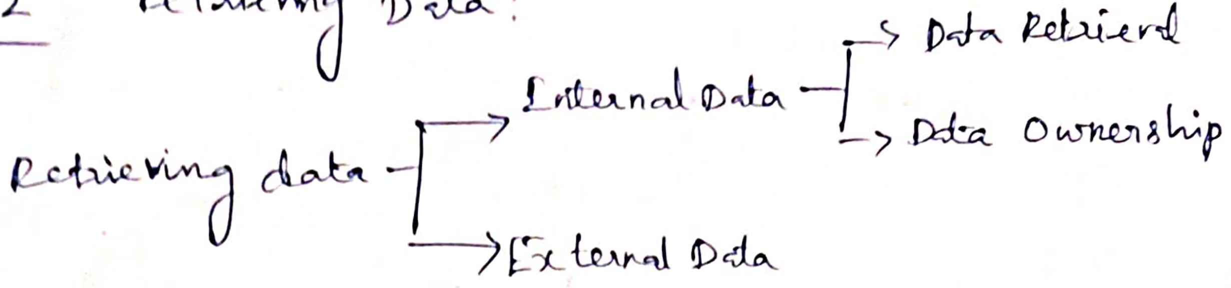
* Out Come should be a Clear research goal
  - → good understanding of the Context.
  - → well defined deliverables
  - – plan of action with a timetable.

Creating a project Charter:

A project Charter Covers the following
- A Clear research goal
- The project mission and Context
- How to do perform the analysis
- what resources are expected to use
- proof of Concepts
- Deliverables and a measure of Success
- A time table.

* A client Can use this information to estimate the project Cost, data and people required for Completion of project.

Step-2 Retaining Data:


Retrieving data
Internal Data
Data Retrieved
Data Ownership
External Data

* 2nd Step of data Science process is to retrieve the required data.

* Many Companies have already Collected & Stored the data for use internally

* More organizations are making high quality data freely available for public & commercial use.

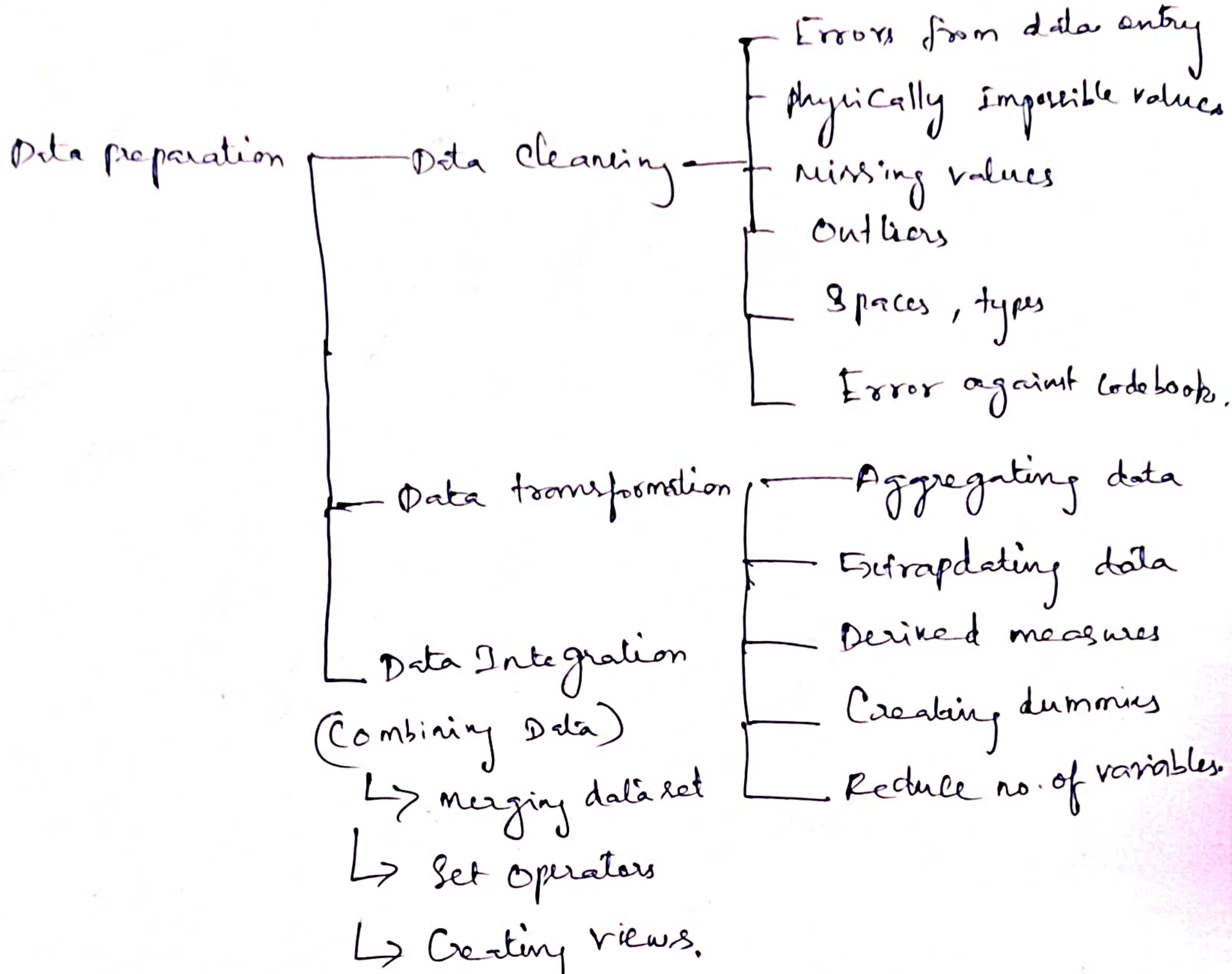
* Data Can be Stored in official data repositories Such as databases, data marts, data warehouse & data lakes.

* Getting access to data is difficult task. So Organization have policies which Controls the access of data by everyone in that organization.
[i.e Access rights - who can access what data]

A List of open data provides

1. Data.gov - the home of US gov. open data
2. Open-data.europa.eu - the home of European Commission's open data.
3. Data.worldbank.org - open data initiative from the world bank.

Step-3: Cleaning, integration and transforming data.



Data preparation

- Data cleaning
  - Errors from data entry
  - Physically impossible values
  - Missing values
  - Outliers
  - Spaces, types
  - Error against codebooks.
- Data transformation
  - Aggregating data
  - Extrapolating data
  - Derived measures
  - Creating dummies
  - Reduce no. of variables.
- Data Integration (Combining Data)
  - ↳ merging data set
  - ↳ Set operators
  - ↳ Creating views.

* The data received from the data retrieval phase is like a "diamond in the rough". So it must be sanitized & prepared for use in modelling & reporting phase.

Possible error's with example :

1) Data entry process - Entry as "Godo" instead of "Good".

2) Redundant white Spaces - Redundant white Spaces at the end of a String Causes errors, very hard to detect.

3) Impossible values - people taller than 3 meters (or) people with an age of 299 years.

4. outliers - is a data that lies abnormally far away from other values in a data set.

Eg : 15, 25, 30, 28, 35, $\boxed{98}$

↓ outlier.

5. Deviations from a Code Books - Detecting errors in large data set against a Code Book.

6. Different units of measurements - Recalculate

Eg. prices/gallon etc.

7. Different levels of aggregation - Bring to Same level of measurement.

Eg: Salary/week

* A good practice is to Correct errors as early as possible.

* Different ways of Combining data Sources.

① joining :- enriching an observation from one table with information from another table.

Eg:

Client	Item	Month
John	Coco-Cola	Jan
James	pepsi	Jan

Client	Region
John	TN
James	AP

Result of joining two tables.

Client	Item	Month	Region
John	Coco-Cola	Jan	TN
James	pepsi	Jan	AP

② Appending Table :- Adding the observation of one table to those of another table.

Eg:

Client	Item	Month
John	Coca-Cola	Jan
James	Pepsi	Jan

Client	Item	Month
Jack	Zero-Cola	Feb
Don	pepsi-Cola	Feb

Client	Item	Month
John	Coca-Cola	Jan
James	Pepsi	Jan
Jack	Zero-Cola	Feb
Don	Pepsi-Cola	Feb

* To avoid duplication of data, one Can virtually Combine data with Views.

* also data enrichment Can be done by adding Calculated information.

Transforming data:

Transforming data makes the data suitable for Data Modeling.

Eg. Transforming x to log x.

- Reducing the no. of variables (principal component Analysis)
- Turning variables into dummies.

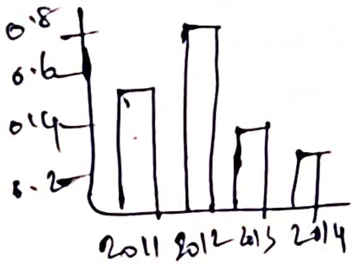
Dummy variables Can only take two values true (1) (or) false (0).

Step-4: Exploratory data analysis:

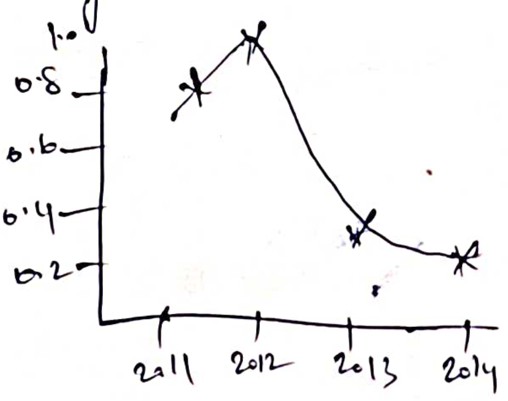
Helps to gain deep understanding of data & the interactions b/w variables.

Data Exploration →
- Simple Graphs
- Combined
- Link & brush
- Non graphical techniques.

* A bar chart, a line chart & histograms used in exploratory analysis.



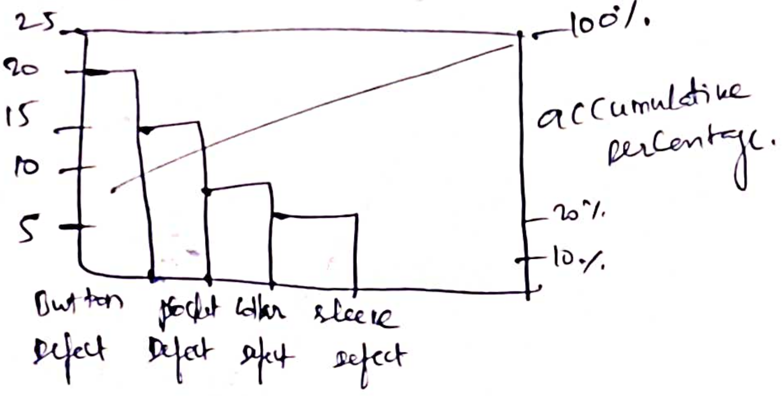
Bar chart



Line chart

Combined graph- pareto diagram.

pareto diagram is a Combination of a bar graph & a line graph.

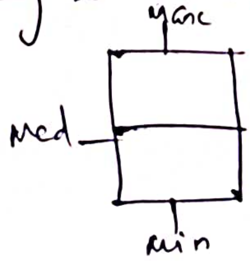


Link and Brush : helps to Combine & link different graphs & tables. so Changes in one graph are automatically transffered to the other graph.

The following Shows the average Score per Country for Question.

Box plot – Shows the maximum, minimum, median & others Characterizing measures.

others are Tabulation, Clustering & other modeling tech.



Build the model:

Building model with the goal of making better predictions, Classifying objects.

The Components of model building

Data Modelling
- Model & variable selection
- Model execution
- Model diagnostic & model Comparison.

* Building a model is an iterative process.

* It depends on statistics & machine language learning techniques.

① Model & variable selection:

* Many modelling techniques are available,

* Choosing the right model based on factors. Such as model performance, project requirements, easy to implement, difficulty in maintenance & easy to explain.

② Model execution:

* Implemention of Choosen model in code.

* Python has libraries such as Stats models (or) scikit - learn to implement models.