Session-1

what is data science :-

A multi-displinary field that uses scientific method, proless & algorithms and systems to extract knowledge & insights from structured & unstructured data.

components of datascience.

Domain Expertise & scientific methods.

- Analysis
- Mathematical & statistical models.
- Scientific tools & methods

Data science.

Technology

- python language
- operating system.
- Application design
- Library.
- data processing tools

Data science in sectors

- Data science is multidisplinary
- In healthcare. etc.

Data scienle applications

→ predection. → classification → Recommendations

→ pattern detection & grouping

→ Anomaly detection.

→ Recognition. → Actionable processes

→ storing & ranking → segmentation.

→ optimization → Forcasts.

Bigdata overview

→ Huge volume of data

→ Complexity of datatypes & structures.

→ Speed of new data creation & growth.

Sources of Bigdata

→ Sensors → Audio

→ social → server / Network log.

→ Text

→ location

→ image

→ Video

Industry:

Big data analytics in industry Verticals

Banking & securities, Communication media & enter tainment, Health care providers, Education, manufacturing & Natural resources, Government, Insurance, Retail & wholesale traders transportation, Energy & utilities.

Session-2

Data analytics lifecycle.

Life cycle overview

→ The data analytics life cycle is designed for big data problems & data science projects.

→ with six phases The project work can occur in several phases simultaneously.

→ The cycle is iterative.

→ work can return to earlier phases as new information is uncovered.

Description of key Roles in analytics project

→ Bussiness user - understands the domain area.

→ Project sponser - provides requirements.

→ project manager - ensures meeting objectives

→ Bussiness Intelligence Analyst - provides business domain expertise based on deep understanding of the data.

→ Database Adminstration (DBA) - creates DB environment.

→ Data engineer - provides technical skills assists data management & extraction, supports analytic sandbox.

→ Data scientist - provides analytic techniques & modeling.

→

Description of key Roles in analytics project

→ Bussiness user - understands the domain area.

→ Project sponser - provides requirements.

→ project manager - ensures meeting objectives

→ Bussiness Intelligence Analyst - provides business domain expertise based on deep understanding of the data.

→ Database Adminstration (DBA) - creates DB environment.

→ Data engineer - provides technical skills assists data management & extraction, supports analytic sandbox.

→ Data scientist - provides analytic techniques & modeling.

→

The Analytics DELTA

Data ..... breadth, integration, quality.

Enterprise ..... approach to managing analytics

Leadership ..... passion & commitment

Targets ..... first deep, Then broad

Analysts ..... professionals V amateurs.

Data analytics lifecycle phases

phase 1: Discovery.

phase 2: Data preparation.

phase 3: model planning.

phase 4: Model building

phase 5: Communicate Results.

phase 6: operationalize.

phase 1: Discovery - Activities

1. Learning the Business Domain
2. Acquiring Resources
3. Framing the problem Definition

4. Identifying key stakeholders.

5. Interviewing the Analytics Sponsor

6. Developing Initial Hypotheses.

7. Identifying potential Data sources.

phase 2: Data preparation.

it has the following activities

1. preparing the Analytic Sandbox.
2. performing ETLT.
3. Learning about the data.
4. Data conditioning.
5. survey & visualize.
6. common Tools for data preparation.

→ 2.1 preparing the analytic Sandbox.

→ 2.2 performing ETLT (Extract, Transform, Load, Transform)

2.3 Learing about the data.

2.4 Data conditioning

2.5 survey & visualize

2.6 Common Tools for data preparation.

Session-3

phase3 model planning: where the team determines the methods, techniques & workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables & the most suitable models.

phase4: model building: The team develops data sets for testing training & production purposes. In this phase the team builds & executes models based on the work done in the model planning phase.

Phase 5: Communicate results:

The team, in collabration with major stake holders, determines if the results of the project are sucess (or) failure based on the Criteria developed in phase 1.

Phase 6: Operationalize:

The team delivers final reports, briefings, code & technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

Tools: model building phase

Commercial Tools:

- → SAS Enterprise Miner
- → Spss modder
- → Matlab
- → Alpine Miner
- → STATISTICA & Mathematica

Free & open source tools:

- SQL
- python,
- WEKA
- octave
- R & PL/R

Benefits & use of data science

Example 1: Google Adsense, which collects data from internet users so relevant commercial messages can be matched to the person browsing the internet.

Example 2: Platform Benefits & uses of data science & big data analytic is another example of real time personalized advertising.

Facets of data

- Structured
- Unstructured.
- Natural language.
- Machine-generated.
- Graph-based.
- Audio, video & images.
- streaming.

1. Structured data

Structured data is data that depends on a data model and resides in a fixed field within a record. As such it's often easy to store structured data in tables within databases (or) Excel files shown in figure. SQL (a) Structured Query language is the preferred way to manage and query data that resides in database.

2. Unstructured data

→ Unstructured data is data that isn't easy to fit into a data model because of the content is content ~~specific~~ specific (or) varying.

→ The thousands of different languages & dialects out there further complicate this. A human written email, as shown in figure. is also a perfect example of natural language data

3. Natural language.

→ Natural language is a special type of unstructured data; it's challenging to process because it requires knowledge of specific data science techniques & linguistics

4. Machine generated data

→ Machine-generated data is information that's automically created by a computer, process, application (or) other machine without human intervention Machine-generated data is becoming a major data resources and will continue to do so.

5. Graph based (or) network data.

→ Graph data can be a confusing term because any data can be shown in graph in this case points to mathematical graph theory.

6. Audio, image & video

→ Audio, image & video are data types

that pose specific challenges to a data scientist tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.

7. Streaming data

→ While streaming data can take almost any of the previous forms, it has an extra property the data flows into the system when an event happens instead of being loaded into a datastore in a batch.

Session-4

Data science process & Data wrangling

→

Data science process

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.

→ Data science is mostly applied in the context of an organization. When the business asks to perform a data science project. This charter contains info such as what is the work going to research.

2. ~~Retrieving data~~

2. Retrieving data

→ The second step is to collect data. stated in the project character which data is need & where can find it.

3. Data preparation

→ Data collection is an error prone proces in this phase enhanie the quality of the data and prepare it for use in subsequent steps.

4. Data exploration

Data exploration is ~~concerned~~ concerned with building a deeper understanding of data. Try to understand how variable interact with each other, the distribution of data, & whether there are outliers.

5. Data modeling (or) model building

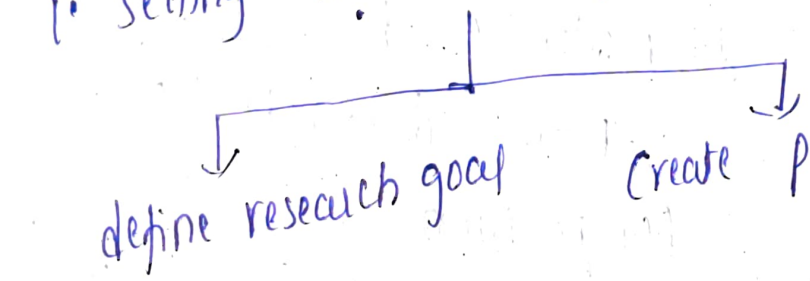
→ In this phase use models, domain knowledge and insights about the data found in previous steps to answer the research question.

6. presentation & automation

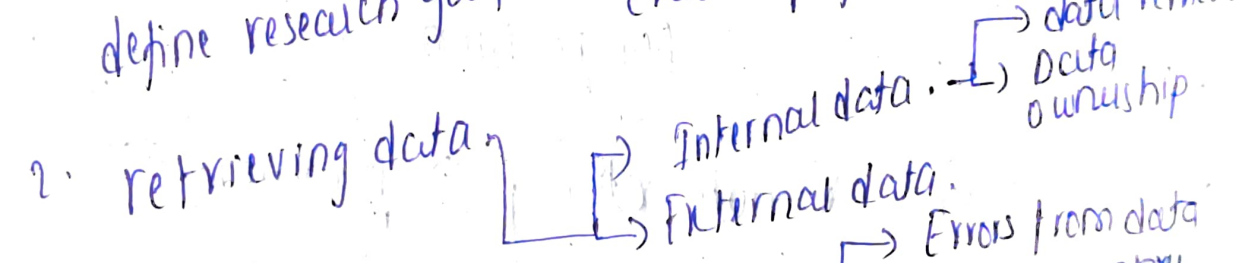
→ Finally, present the results to business. These results can take many forms, ranging from presentations to research reports.

These six are again sub divided..

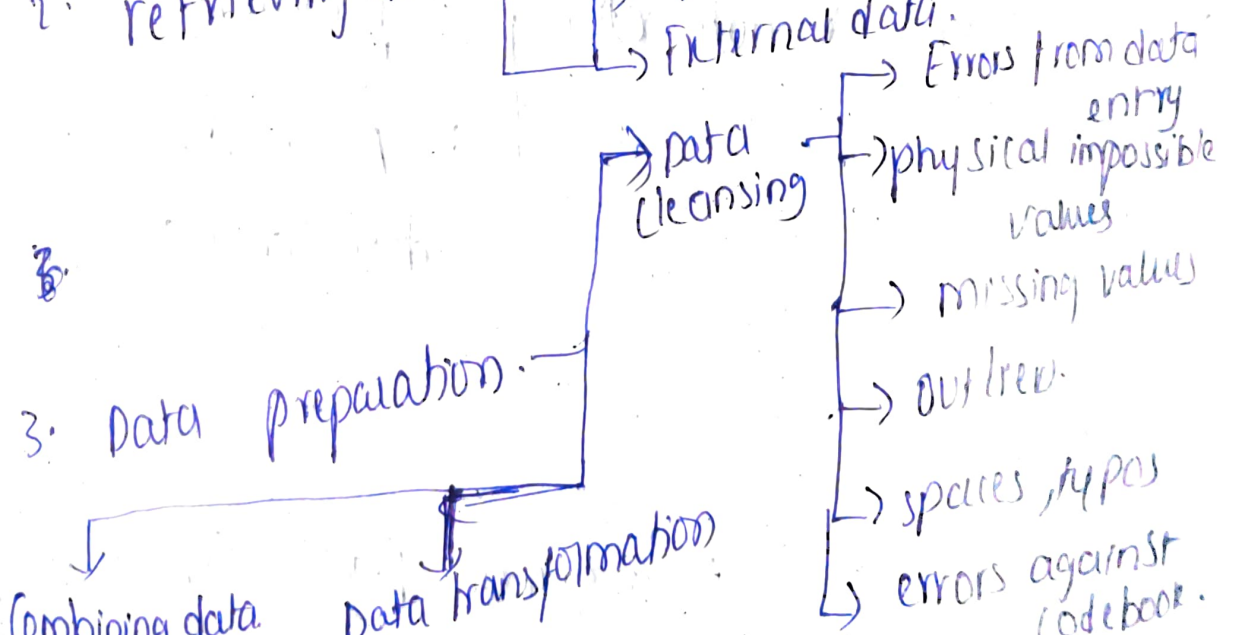
1. setting the research goal.



2. retrieving data



3. Data preparation.



4. Data exploration
   - Simple graphs
   - Combined graphs
   - link & brush
   - Non graphical techniques

5. Data modeling
   - model and variable selection
   - model execution.
   - model diagnostic & model comparision.

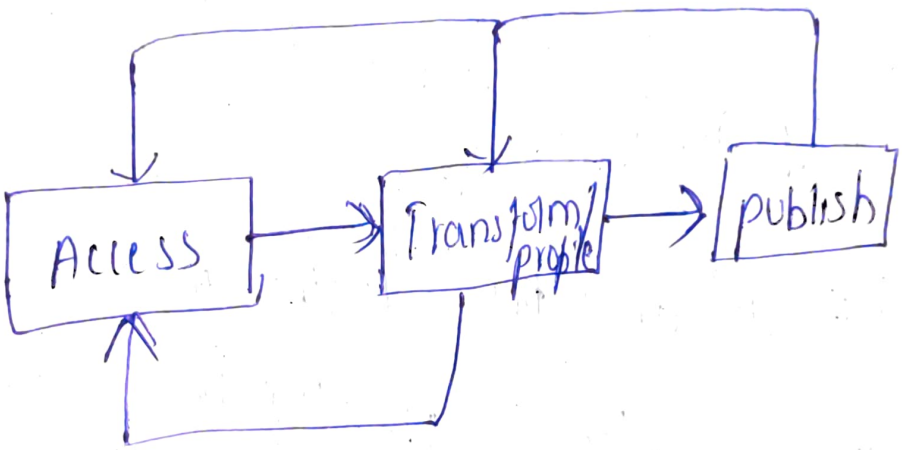
6. presentation & automation
   - presenting data
   - Automating data analysis.

Data wrangling

→ Data wrangling is a generic phase capturing the range of the tasks involved in prepaing the data for analysis. it begins with accessing data, ~~sometimes~~ sometimes, access

is gated on getting appropriate permission & making the corresponding changes in data infrastructure. Access also involves manipulating the locations & relationships between dataset.

A simple diagram



Session 5

Data analytic process.

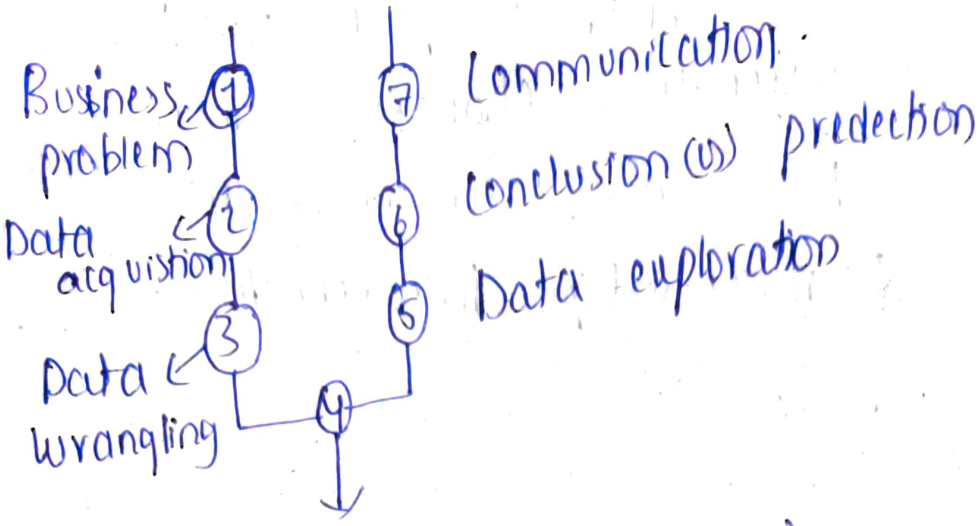
Why data analytics?

Data by itself is just an information source But unless you understand it, you will not be able to use it effectively.

Introduction

* Data analytics:

it is a combination of process to extract data from datasets.



1. Business problem

The process of analytics begins with questions

(or) business problems of stakeholders.

2. Data acquisition

Data scientists Expertise

→ File handling

→ File formats

→ web scrapping.

(3) Data wrangling

Data Cleansing

Data manipulation

Data wrangling.

4. exploratory Data analysis (EDA)

→ EDA approach studies the data to recommend suitable models that best fit the data.

→ The focus is on data : its structure, outliers, and models suggested by the data.

Significance of Eda, steps in eda, eda mawing sense of data, eda: measurement scales, eda quantitative technique, eda graphical techniqu all in pr (course material 1/2).

(5) Data Exploration : model selection.

model selection

- Based on the overall data analysis process.
  - → should be accurate to avoid iterations
  - → Depends on pattern identification & algorithms.
  - → Depends on pattern identification & algorithm
  - → Depends on hypothesis building & testing

7. Communication

Forms of data analysis presentations:

- → visual graphs
- → plotting maps.
- → Reports
- → white paper reports.
- → power point presentations

Session-6

EDA: Graphical Techniques

Shape base EDA : skewness

→ The literal meaning of "skew" is a bias, dragging (or) distortion toward some particular value, group, subjects (or) direction.

→ A measure of skewness is a numeric metric to concisely summarize the degree of asymmetry of a unimodal distribution that can be compared with other similar numbers.

Calculating Skewness

$$g_1 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2 / 3}{s^3}$$

where $\bar{y}$ is the mean of

s is the standard deviation

N is the number of datapoints.

The above formula of for skewness is refered to the Fisher-Pearson coefficient of skewness

Skewed graphs in pg:24.

Shape based EDA : Kurtosis

→ kurtosis is a statistical measure that defines how heavily the tails of distribution differ from the tails of normal distribution.

→ In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.

Types of kurtosis
- → mesokurtic
- → platykurtic
- → leptokurtic

Mesokutric

→ Data that follows a meso kutric distribution shows an excess kutrosis of zero (or) close to zero.

Lepto kutric

→ Lepto kutric indicates a positive excess kutrosis

Platy kutric

→ A platykutric distribution shows a negitive excess kurtosis.

Histograms

→ Histogram graphically Summerizes the distribution of univariate dataset. It shows:

→ The centre (or) location of data.

→ The Skewness of data.

→ The presence of outliers.

→ The presence of multiple modes in the data.

Scatter plots

→ A scatterplot represents relationships between two variables.

CO2

Data visulization

Session →

Data analysis

It is the most crucial steps that deal with descriptive statistics. and analysis of the data.

Data analysis techniques

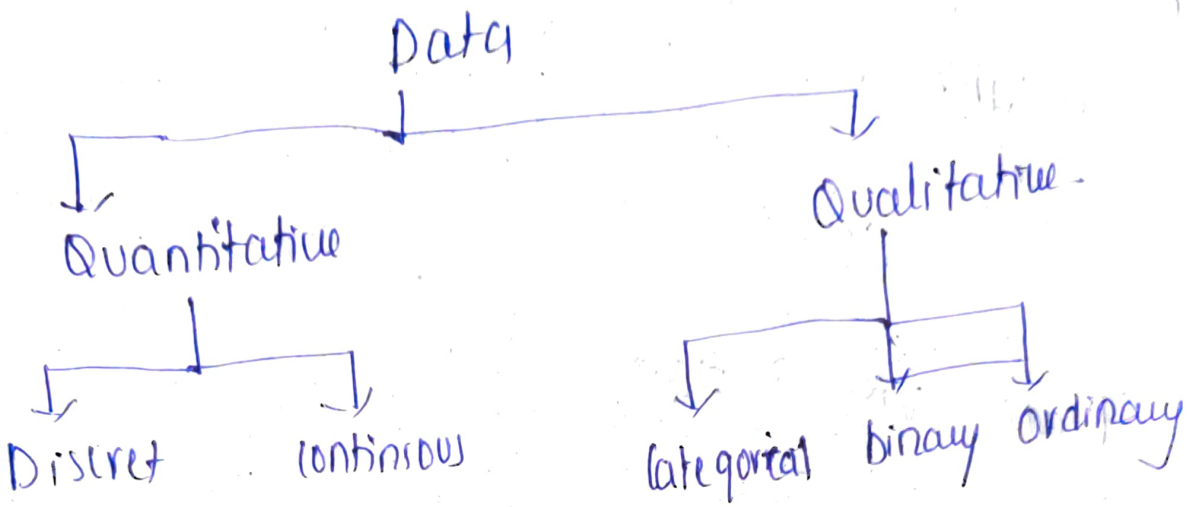
→ Data summarizations are summary tables

→ Graphs.

→ Descriptive statistics, inferential statistics, correlation statistics

→ Searching, grouping & mathematical models

Data types for plotting



Quantitative data

The information is recorded as numbers. it represents an objective measurement (or) a count

continious data

continious variables can take any numeric value. it can meaningfully divided into smaller increments including fractional & decimal values.

Discrete Data

Discrete data has a limited number of possible values.