27/01/25

UNIT-III

DATA SCIENCE (DS)

Data, information, knowledge.

Visualisation of Data, characteristics of DS

Data classification:- Data → 3 ways

1. Structured 2. Unstructured 3. Semi-structured.

Individual wells:-

1. Ecological well 2. Cultural well 3. HR (Human Resources) 4. Marketing well 5. Finance well 6. Online Transaction Processing data well 7.

Feature Selection of Data Sciences

* It is the process of selecting a subset of relevant features for use in model construction.
* It is also called variable selection or attribute selection
* It reduces the complexity of model
* It either improve or maintain accuracy of model.

Data Science:-

Data science is an interdisciplinary field that use algorithms, procedures, process to examine large amount of Data in order to uncovered hidden patterns. It generates insights & direct decision making

(or)

Data Science is a study of extracting meaningful insights for business problems & decision making in the field of interdisciplinary areas.

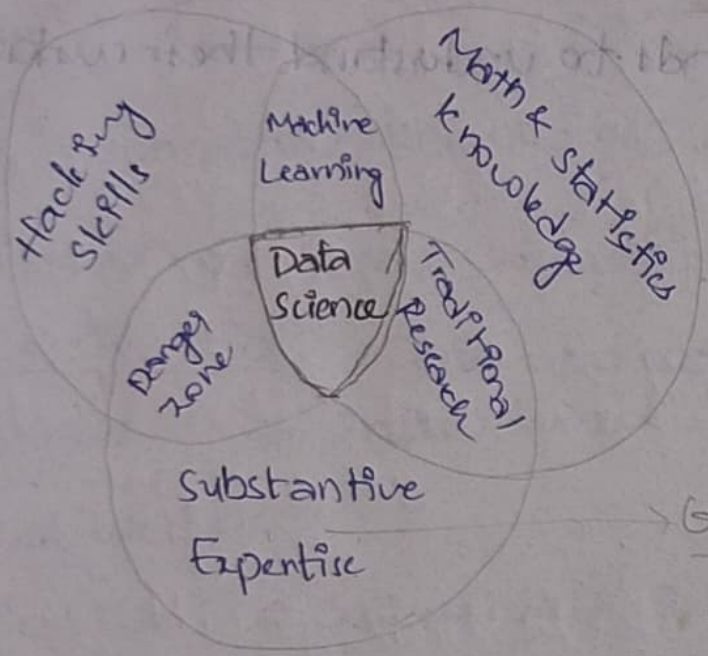
Importance of Data Science:-

* Data science helps brands to understand their customers in a a much enhanced &

By the use of these results, we can analyze as a solution for the current situation as well as Decision making.

The Landscape of Data science:-

Diagram(p.To)



→ Good at anywhere

As an organisational discussion, the actual process is going on 'industry is to look online & see what current discussions are taking place. This is not neccessary to till us what other people are thinking, what data science is doing, what are the latest applications are pursevivasive, such type of several questionaires come across in your mind. To overcome this, "drew" conveys Venn diagram of representation about the data science, in 2010.

He also mentioned the advanced skills of data takes from "raise of the Datascience" which includes statistical like traditional analysis of data, data mining by using, parsing, scrapping & formatting. The Above Venn diagram says that Data Science is an integration path of hacking skills, mathematical & statistical knowledge, substantive expertise in the field of technical skills. All these 3 features are intallocating with Data science & its applications

Bigdata and Datascience

The Terms "Bigdata" & "Data Science" often emerge as pivotal concepts driving innovation & decision-making. Despite their frequent interchangeability in casual conversation, Bigdata & Data science represent distinct but interelated fields.
understanding

Big data v/s Data Sciences

Big data consists of 3 v's volume, velocity, variety.

Volume: Big Data involves large datasets that are too complex for traditional data processing tools to handle. These datasets can range from terabytes to petabytes of information.

velocity: Bigdata is generated in real-time or near real-time requiring fast processing

variety: Gui & XML files

Data Scientists use their Expertise (Data scientist activities)
Analyze, Model, Interpret

modul: using statistical Models, ML algorithms, they create predictive models that can forecast

Differences:-

Without data there is no data science

Aspect	Bigdata	Datascience
Definition	Handling & processing vast amt of data	Extracting insights & knowledge from data

Objective	Efficient storage, processing & mgmt of data	Analyzing the data to information, decisions & predict the trends.
Focus	Volume, velocity & variety of data	Analytical & mathematical model & algorithms.
primary tasks	collection, storage & processing of data	Data Analysis, modling & interpretation.
Tools/ Technologies	Hadoop, spark, NOSQL, databases (eg. MongoDB)	python, R programming, Scikit-learning. Tensorflow
Data Types	structured, unstructured & semi-structured data	Processing & clean data for analysis.
Outcomes	Accessible data repositories for analysis	Actionable insights predictive model.
Skillset	Data engineering, distributed computing	Data scientists, Machine learning. Engineering.
Typical roles	Data Engineer, Big Data Analysis	Statistical analysis, machine Learning programming. etc
Application	Real time data processing large-scale data storage	predictive analysis, data driven decision making
Key Techniques	Distributed computing & data warehousing	Statistical modeling, ML algorithms.

Advantages & Dis-Adv

Big data

Adv

Able to handle the process & handle large amt of data that cannot be easily managed with Traditional DBMS

Provides a platform for Advance analysis & ML application.

Enable Organisation to gain insights & make data driven decisions based on large amt of data.

Offers potential for significant cost saving through efficient data mgmt & analysis

Dis-Adv

Requires specialized skills & expertise in Data Engineering, Data mgmt, Big data tools & technologies.

Can be expensive to implement & maintain due to the need for specialized infrastructure & s/w

May face privacy & security concerns when handling sensitive data

Data Science

Adv

Provides a framework for extracting insight & knowledge from data to statistical Analysis, Machine Learning & Data visualization techniques.

offers a wide range of applications in various fields such as finance, healthcare, banking & marketing etc.

Helps the organisation, make information decision by extracting meaningful insights from data.

offers potential for significant cost saving through efficient data mgmt & analysis

Dis-Adv.

Requires specialized skills & expertise in statistical & Mathematical Analysis, ML algorithms Analysis, Data Visualization technique

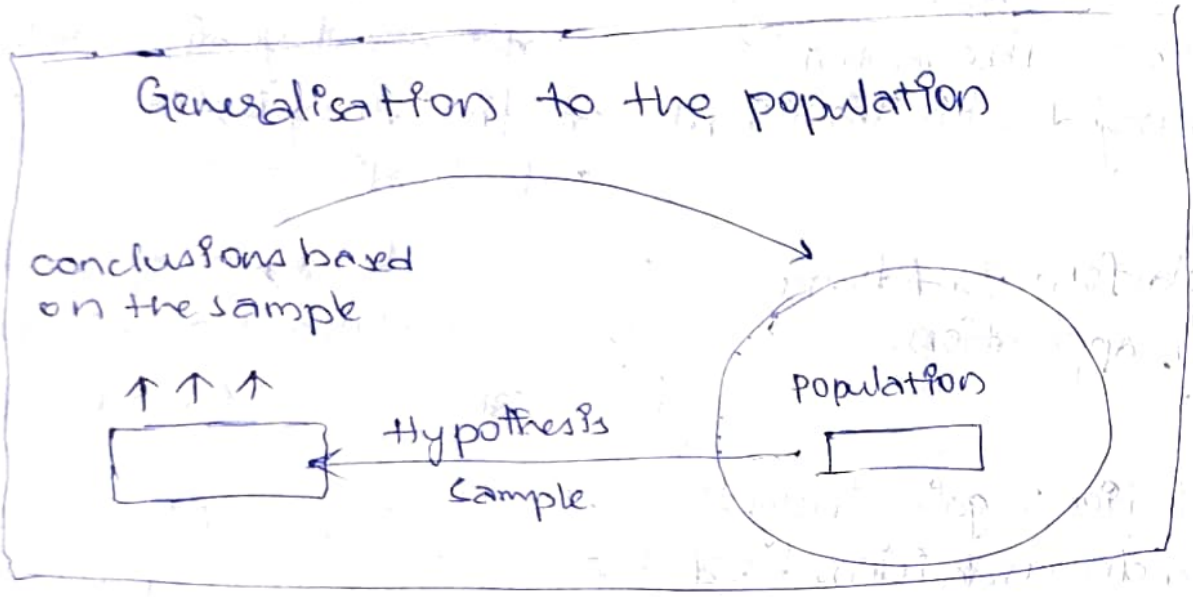
can be time consuming & resource intensive due to the need of data cleaning & processing

May face ethical concerns while dealing with the sensitive data

can be challenging to integrate with the existing systems & process

can be challenging to integrate with the existing systems & process.

Population Inference in data science.



In statistics we generally want to study population. We can think of a population as an entire collection of persons, things or objects under the study. To study the larger population we select a sample. The idea of sampling is select a portion or subset of the larger population & study the population portion to gain the information about the population. Data is the result of a sampling from a population.

Because it takes a lot of time & money to examine entire population, sampling is very pratical technique. If you wish to compute overall grade points average at your school. It would make sense to select sample of students grade point average. Especially this types of surveys may be implemented especially this types of surveys may be implemented especially before precidential elections. opinion pole, samples of 1000 to 2000 people are taken. The opinion pole supposed to represent the view of the people

in the entire country. In each & every promotional activities especially in the manufacturing & marketing sectors. This types of samples will be taken as a business consideration.

Confidence Intervals:-

We often use sample data to estimate population quantities due to the Randomness inherit to sampling & observe sample statistics is almost certainly not equal to the true population Parameters. To qualify the variability surround the sample statistics we can compute a confidence interval (0-10) which provids lower & upper bound for where we think true population values lies. Note that unless we take a sample which consists of the entire population (census), we will never know the true population parameters with absolute certainty.

Branches of statistical Inferences:-

Based on the utilization of data, visualization of data analysis the statistical inference divided into 2 branches.

① Parameter Estimation ② Hypothesis Testing

Parameter Estimation:-

This is another primary goal of statistical inference. Parameters are capable of being deduced. They are quantified traits or properties related to the population you are studying. This estimation if not an impossible task most of the time we can use this estimation. This parameters estimation can again there are 2 broad main areas.

1. Point Based Estimation 2. Interval Based Estimation.

Hypothesis Testing:-

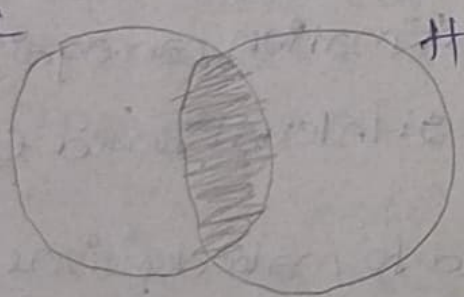
Hypothesis testing used to make decisions or draw conclusions about a population based on samples. It involves formulating a hypothesis about the population parameter, collection sample data & then using statistical method to determine whether the data provid enough evidence to reject or fail to reject the hypothesis

An important components of hypothesis testing is coming from inference which allows us to analyze the evidence provided by the sample to access some claim about the population. Here you can discuss about.

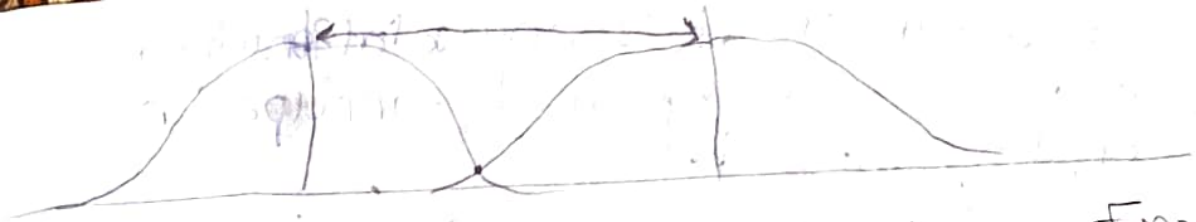
a. One sample hypothesis
b. Two sample hypothesis

One sample hypothesis Test compared single parameter to specified value.

A Two sample hypothesis test access the equality of parameters from a different populations. For example, you are working with a reputed company, your house is near by your company. Here, house & company are 2 samples as a profession, you're giving prominance to house as well as company. Here, house is treated as H & company is treated as C. Both the areas intersect part of these 2 samples are considered as a denced population sample. Here, the main of 2 samples. 2 different random samples with a different population atleast up to some extent, you need to test against these 2 samples with different variations such as



Sometimes, it is called as Paired-T-tested



$$t = \frac{(\mu \text{ sample1} - \mu \text{ sample2}) - (\mu_1 - \mu_2)}{s_p}$$

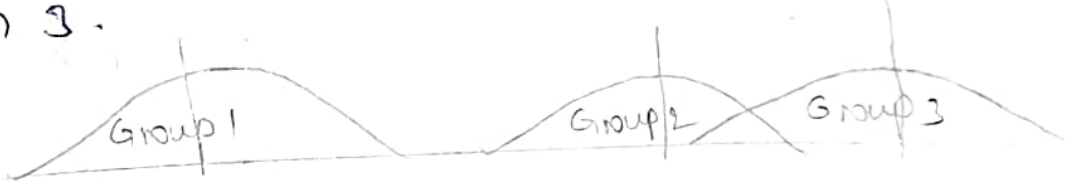
$$F_{max} = \frac{s^2 (\text{largest})}{s^2 (\text{smallest})}$$

$$s^2 = \frac{SS}{df}$$

Hypothesis testing more than 2 samples:

ANOVA:- Analysis of Variance

If we want to compare the means of more than 2 groups 1 procedure is available which is called Analysis of Variance (ANOVA). The name seems to be strange, because we are comparing mean. The word come from the fact that this procedures makes relatively strong assumption but the availability of each group, we are comparing is the same. A rule of thumb when using ANOVA with the ratio of largest S.D of the group with the smallest S.D should be not more than 3.



chi square testing:-

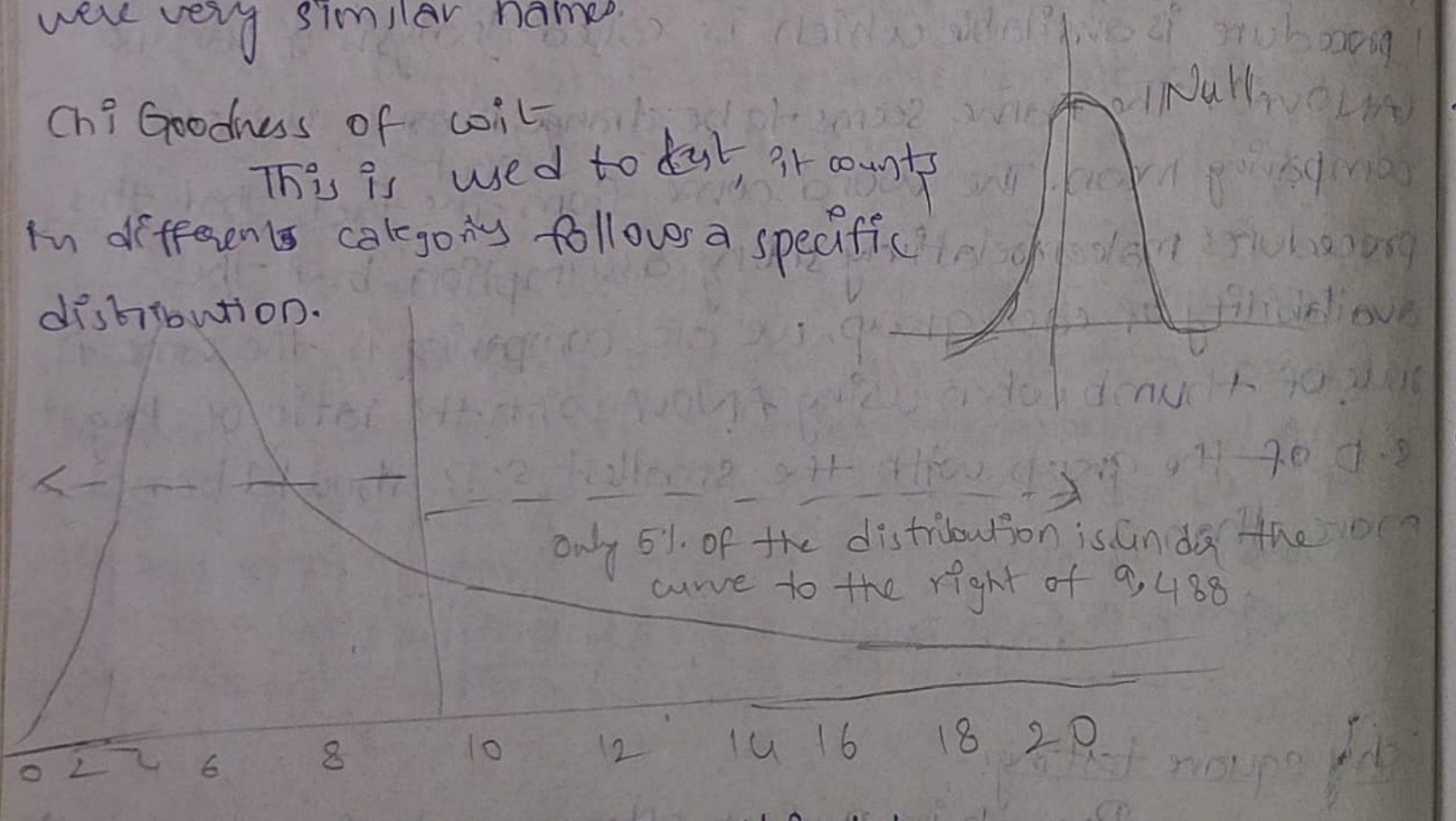
Research in Business often generate frequency data. This is certainly the case in biggest most opinion surveys in which the person interview is ask to respond the question by making say "Agree," "Not Agree" or "disagree" or some other collection of categories.

In this case, the investigator might be concern with a determining what proportion of relationship mark or response

each of the choice or whether there is any relationship b/w the opinion marked & such as age, gender, occupation, religion & soon of the respondent.

Chi square method make posible for meaningful analysis for frequency data by permutting the the comparison & frequency actually observed of the with the frequencies which would be expected with a null hypothesis for true. At first glance of chi-square test procedure can be confusing as there are 2 different tests were very similar names.

Chi Goodness of wilt
This is used to test, it counts in different categories follows a specific distribution.



Suppose by illustrating this below example Bar galaxy chocolate which want to determine if customers have a preference for any of the following candy Bars. The random sample of 200 people, it was found that
1. 43 prefered the frosty bar.
2. 53 prefered the Galaxy milk chocolate.

3. 60 prefered galaxy spl dark chocolate
4. 44 prefered munchy spark.

For the Goodness of fit test, the null hypothesis states that customers have no preferences any of the 4 candy Bars (1, 2, 3, 4). This is all four candy bars are equally prefered. The Alternative hypothesis states that the preference probabilities are not all the same.

Statistical Modelling:

Statistical Modelling is like a formal depiction of a theory, it is typically described as the mathmatical relationship blw random & non-random variables. The science of statistics is the study of how to learn from data. It helps you to correct the right data, perform the correct analysis & statistical knowledge, statistical modling is key to making discoveries, data driven decisions & predictions.

statistical modelling helps you to diffitiate blw reasonable & debious conclusions based on ~~con~~ quantitative ~~units/realities~~ evidence. Analysis & predictions made by statisticians are highly ~~worthy~~ trust coorthy. A statistician can help investigators avoid various analytical traps along the way.

Statistical Modeling Techniques:-

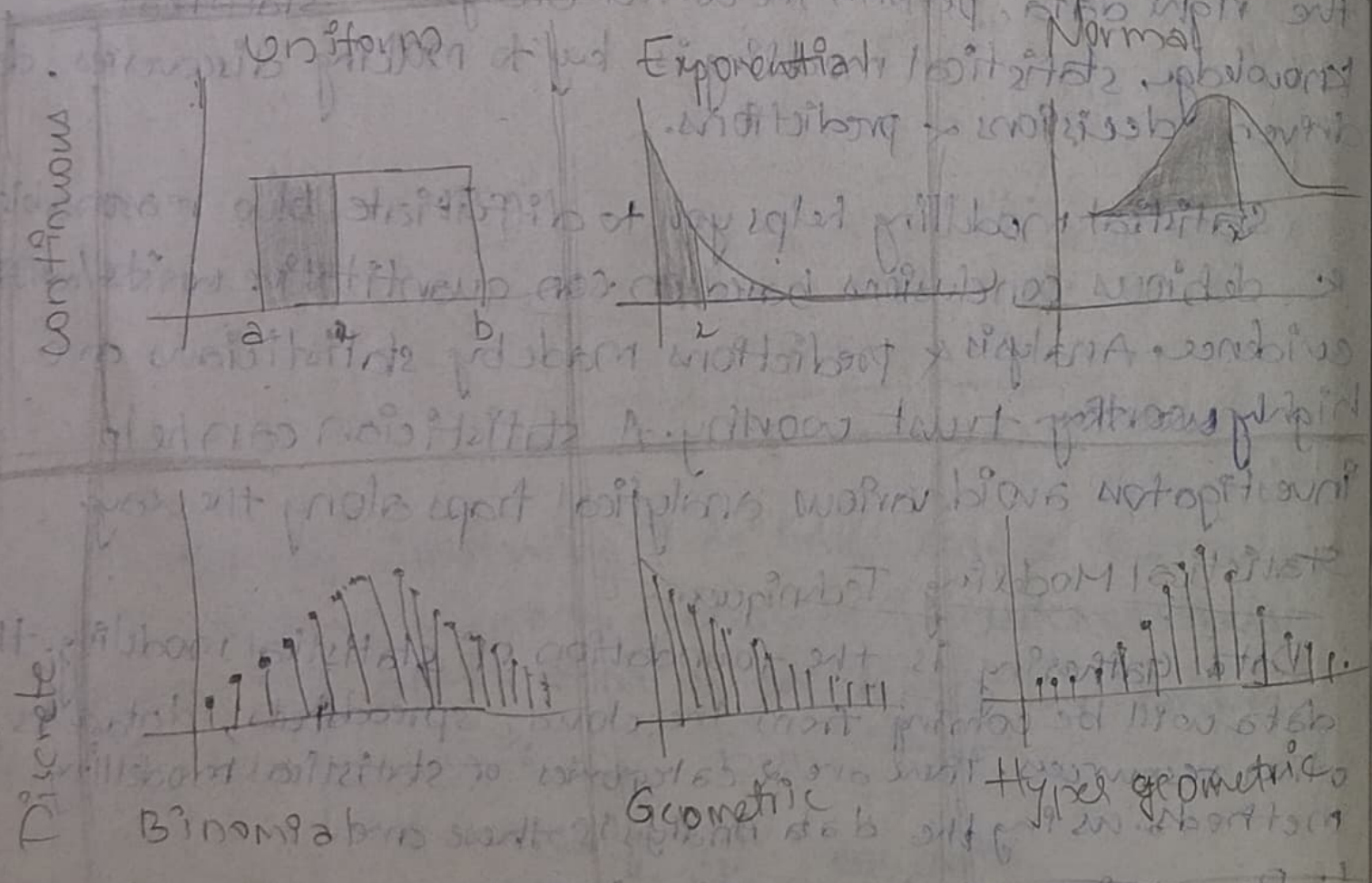
Data Gathering is the foundation of statistical modeling the data will be coming from the cloud, spreadsheets, datasets, other resources. There are 2 categories of statistical modelling methods. using the data Analysis these are

1. Regression Model, 2. Classification Model

1. Regression Model:- A predictive model that analyse the independent - Dependent variable. the most common regression model are, logistical, polygonial, & Linear. These modules determine information, variables forcasting & modelling

Classification Model:- An algorithm analyzes & classifies a large & complex set of data & its great points, In this there are several models includes decision based, Navy based, the nearest neighbourhood, Random forcast Random forest, neural networks, storage area snapshot.

Probability distribution:



We will divide this probability distribution whether the data is discrete or continuous.

No matter what field you are in, probability statistics will be there. Economics, finance, trading, social science, natural science & of course data science is indispensable The significance of Data science is about understand the behaviour & properties of the variables. And this is not possible without knowing what distribution they belong to, simply put the probability distribution is a way to represent possible values a variable may be take & their respective probability.

Continuous Probability Distribution:-

The continous probability Distribution focus on

1. Uniform distribution (continuous) or normal
2. Exponential distribution
3. Normal distribution
4. Chi square distribution
5. Log-normal distribution.

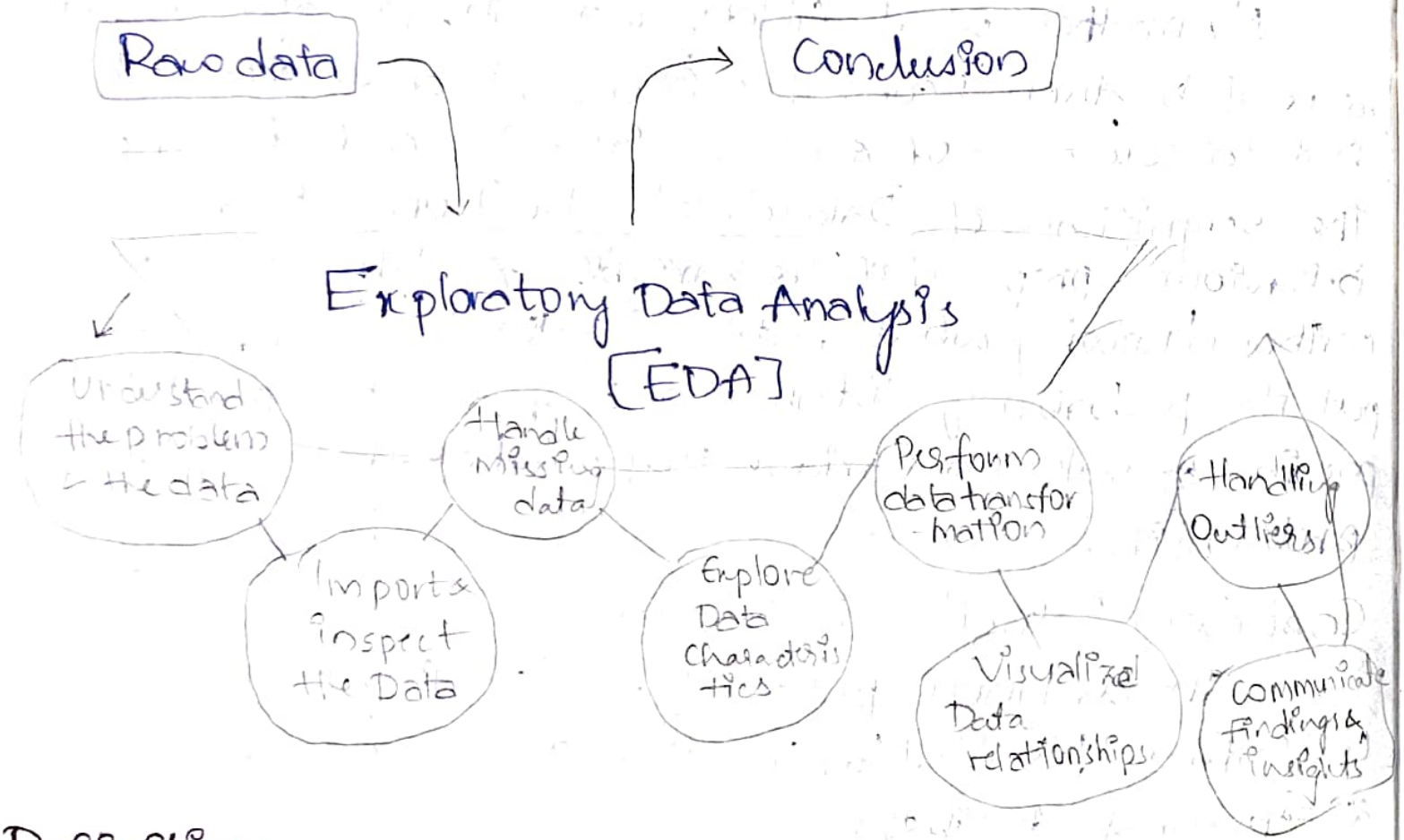
And und distribution again we have several types of distribution. such as

* Binomial distribution
* Geometric distribution
* HyperGeometric distribution
* Berno-Li distribution.
* Poison distribution

11/02/25

UNIT-2

Exploratory Data Analysis & its tools



Definition:

Exploratory Data Analysis (EDA) is a data Analysis process that aims to understand the data in depth & learn its different characteristics. Often using visual means, this allows one to get better free for the data & find useful patterns

It is crucial to understand in depth before you perform data analysis & run your data to an algorithm. You need to know the patterns in the data & determine which variables are important, unimportant in the data

outputs. Further, some variables may have correlation with other variables, you also mean to recognize errors in the data.

Ex:- For example, EDA can do all of these, it helps you gather insights better sense the data & remove irregularity & unnecessary values.

Take a Student Marks statement as an example for a particular class by using EDA, we can analyze the entire class vizualize as pie or bar graph diagram.

By the above diagram, we can easily perform analysis & decision making especially data science application.

Exploratory data Analysis (EDA) all of these, it helps you to prepare data set of analysis. It Allows you to a meaningful Machine Learning model to predict our dataset better.

- It gives you more accurate result
- It also helps us to choose a better Machine Learning model.

Steps in EDA (Importance of EDA Activities)

1. Understand the Data:-

Familiarize yourself with the data set, understand the domain, & identify the objectives of the analysis

2. Data Collection:

Collect the required data from various sources such as databases, web scraping, or APIs.

3. Data Cleaning:-

* Handle missing values: Impute or remove missing data
* Remove duplicates: Ensures there are no duplicate records
* Correct data types: Convert datatypes to appropriate formats
* Fix errors: Address any inconsistences or errors in data

4. Data Transformation:
* Normalize or standardize the data if necessary
* Create new features through feature engineering
* Aggregate or disaggregate data based on analysis needs

5. Data Integration.
Integrate data from various sources to create a complete data set.

6. Data Exploration:-
- Univariate Analysis:- Analyze individual variables using summary statistics & visualizations (Eg:- histograms, boxplots)
- Bivariate Analysis:- Analyze the relationship b/w 2 variable with scatterplots, correlation, coefficients & cross-tabulations.
- Multivariate Analysis:- Investigate interactions b/w multiple variables using pair plots & correlation matrices.

7. Data Visualization:-
Visualize data distributions & relationships using visual tools such as bar graphs/charts, line charts, scatterplots, heatmaps & boxplots.

8. Descriptive Statistics:-
calculate central tendency measures (mean, median, mode) & dispersion measures (range, variance, standard deviation)

9. Identify Patterns & Outliers
Detect patterns, trends & outliers in the data using visualizations & statistical methods

10. Hypothesis Testing:-
Formulate & test hypotheses using statistical tests (eg t-tests, chi square tests) to validate assumptions or relations in the data

11. Data Summarizations:-
Summarize findings with descriptive statistics, visualizations & key insights.

12. Documentation & Reporting:
Document the EDA process, findings & insights clearly & structured
Create reports & presentations to convey results to stakeholders.

13 Iterate & Refine:
Continously refine the analysis based on feedback & additional questions during the process.

The Role
~~Importance~~ of EDA in Data Science

Tools Required for EDA:-

1. R :- An Open-source programming language & free s/w environment for statistical computing & graphics supported by the R foundation for statistical computing. The R programming language is coidely used among statistics in developing statistical observations & data analysis

2. Python: An interpreted Object-oriented programming language with dynamic semantics. Its high-level, built-in data structures combined with dynamic binding, make it very attractive for RAD, also as to be used as a scripting or glue language to attach existing components together. Python and EDA are often used together to spot missing values in the data set, which is vital so you'll decide the way to handle missing values for machine Learning.

Apart from these functions described above, EDA can Use

Perform k-means clustering: Perform k-means clustering, it's an unsupervised learning algorithm where the info points are assigned to clusters, also referred to as k-groups, k-means clustering is usually utilized in market segmentation, image compression & pattern recognition

- EDA is often utilized in predictive models like linear regression, where its won't to predict outcomes.
- It is also utilized in univariate, bivariate & multivariate.

Philosophy of Exploratory Data Analysis:-

EDA is a philosophy & an approach towards understanding & interpreting data. It is a critical first step in

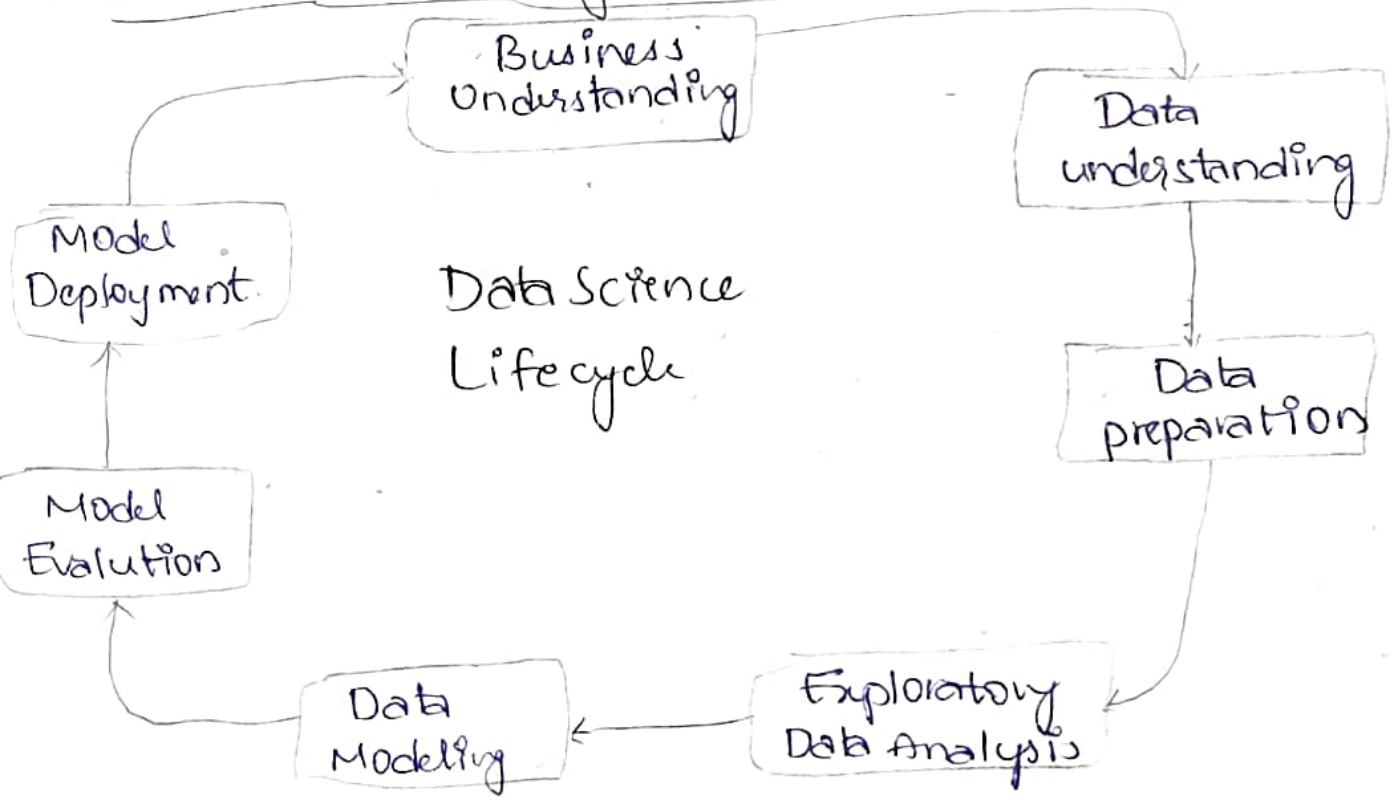
the world of Data science & analytics that allows analyst & researchers to make sense of complex data & to identify patterns, spot anomalies or test hypotheses

By the above below example, we can clearly understand philosophy of EDA.

While at Google, Rachen was fortunate to work along with former Bell labs or AT & T Laboratories. Rachen working for 20 yrs by using mathematical & statistical Analyst also he is good at vizualization techniques. At present in Google mr. ~~Richard~~ Rachen during all EDA Activities as much as ~~sam~~ simple the can explore insights with the unique models, he vizualize understandable patterns when an any other can't compare with his innovative experience.

Mr. Rachen got capertise in the field of EDA, he can change & modify the results & insights for better understanding. Likewise, EDA a great philosophical theory for upcoming Analyst

Data Science Life cycle ~~Activities~~:-



Data Science case Studies in Retail. (Walmart).

Steps involved

1. Prepare of minicase.
2. Find out the problem
3. Developing of Alternative solutions
4. select the best solution among alternative
5. Implement the best solution
6. conclusion

Case study-2: Real Time Pricing Strategy (Uber)

Challenge: Uber needed to adjust its pricing dynamically to reflect real-time demand & supply variations across different locations & times, aiming to optimize driver incentives & customer satisfaction without mannual intervention.

Solution:- Uber introduced a dynamic pricing model called "surge pricing". This system uses data science to automatically calculate fares in real time based on current demand & supply data. The model incorporates traffic conditions, weather forecasts, & local events to adjust prices appropriately.

Overall Impact:-

1. Optimized Ride availability: The model reduced customer wait times by incentivizing more divers to be available during high-demand periods.
2. Increased Driver Earnings: Drivers benefitted from higher earnings during surge periods aligning their incentives with customer demand.

Key Takeaways:-

1. Efficient Balance of Supply & Demand: Dynamic pricing matches ride availability with customer needs
2. Importance of Real-Time Data Processing: The real-time processing of data is crucial for responsive & adoptive service delivery.

Uber's implementation surge pricing illustrates the power of using real-time data analysis to create a flexible & responsive pricing system that benefits both consumers

& service provides, enhancing overall service efficiency & satisfaction.

Data Science Models:-

Statistical Models: -There are several types of Statistical Models especially in Data Science, to understand widely by using several models such as linear

1. Linear Regression models
2. Logistic Regression models
3. Time series models
4. survival models
5. Decision Tree models
6. Neural N/w models
7. Ensemble Models
8. Multivariate Analysis

Introduction to R programming

- It's free
- It runs on variety of platforms including Windows, Unix & MacOS
- It provides an unparalleled platform for programming new statistical methods in an easy & straightforward manner
- It contains advanced statistical routines not yet available in other packages
- It has state-of-the-art graphics capabilities

First, while there are many introductory tutorials (covering data types, basic commands, the interface) none alone are comprehensive. In part, this is because much of the advanced functionality of R comes from hundreds of user contributed packages. Hunting for what you want can be time consuming. & it can be hard to get a clear

The second reason is more transient. As users of statistical

packages, we tend to run one controlled procedure for each type of analysis. Think of PROC GLM in SAS. We can carefully set up the run with all the parameters & options that we need. When we run the procedure, the resulting output may be a 100 pages long. We then shift through this output pulling out what we need & discarding the rest.

R is an open-source programming language used statistical sw & data analysis tools. It is an important tool for Data science. It is a highly popular & is the first choice of many statisticians & data scientists.

- R includes powerful tools for creating aesthetic & insightful visualizations
- Facilitates data extraction, transformation, & loading with interfaces for SQL, spreadsheets & more.
- Provides essential packages for cleaning & transforming data.
- Enables the application of ML algorithms to predict future events
- Supports analysis of unstructured data through NoSQL database interfaces.

Syntax & Variables in R

In R, we use the <- operator to assign values to variables, though = is also commonly used. You can also add comments in your code to explain what's happening, using the symbol '#'.

Ex:1

```
x <- 5 # Assigns the value 5 to x
y <- 3 # Assigns the value 3 to y
```

```
sum result <- x + y
product result <- x * y
print(paste('Sum of x & y :', sum result))
print(paste('Product of x & y :', product result))
```

Output

```
[1] 'Sum of x & y: 8'
[1] "Product of x & y: 15"
```

Design a s/w procedure for sum of 2 numbers, product of 2 Numbers using Java

Ex:-2

```
import java.util.Scanner;
class Sum {
public static void main (String args[]) {
Scanner sc = new Scanner (System.in);
S.O.pln("Enter the values of x & y");
int x = sc.nextInt();
int y = sc.nextInt();
S.O.pln("Sum of x & y :" +(x+y));
S.O.pln("Product of x & y :" +(x*y));
}
}
```

Output

```
Enter the values of x & y
10
20
Sum of x & y : 30
Product of x & y : 200
```

Datatypes & Structures in R.

In R, data is stored in various structures, such as vectors, matrices, lists & data frames.

1. **Vectors:-** Vectors are like simple arrays that hold multiple values of the same type. You can create a vector using the c() function

```
vector <- c(1,2,3,5)
print(vector)
```

Output:-

```
[1] 1 2 3 5
```

2. **Matrices:-** Matrices are 2-dimensional arrays where each element has the same data type. You create a matrix using the matrix() function:

```
matrix_data <- matrix(1:9, nrow=3, ncol=3)
print(matrix_data)
```

Output

```
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

Data Manipulation with R programming:

R libraries are effective for data manipulation, enabling analysis to clean, transform, & summarize datasets efficiently.

Using dplyr for Data Manipulation:

The dplyr package provides a set of functions that make it easy to manipulate data frames in a clean & readable

manner. Some of the key functions in dplyr include
* filter()
* select()
* mutate()
* arrange()
* summarize()

R Console Interface:-

Start the R system, the main window (RGui) with a subwindow (RConsole) will appear.

In the 'Console' window the cursor is waiting for you to type in some R commands.

R Warning:
R is a case sensitive language.
Ex:- FOO, Foo & foo are 3 different objects

R. Datasets:
R comes with a no. of sample datasets that you can experiment with. Type >data()
to see the available data sets. The results will depend on which packages you have loaded. Type
help (datasetname) for details on a sample dataset.

R packages:-

In R language the system is allowed to write user defined functions, packages are called R packages which is in the R Library.
Ex:- There are different packages such as portfolio mgmt

optimization, drawing packages, exporting objects from one interface to another, access objects from different interfaces such as SQL, MS Excel, html & so on.

R Language Applications:

By using R Programming, there are several types of scientific & mathematical analysis as well as statistical analysis can be made for better understanding using of insightation, we can take qualitative decisions & effective visualization

Some of the applications are as follows.

1. Statistical Analysis & Data Visualization.
2. Data Exploration & cleaning
3. Predictive Modeling & Machine Learning
4. Biostatistics & Healthcare.
5. finance & Risk mgmt
6. Social science & Market research
7. Environmental science & climate research.
8.

1. Define Data science & its importance
2. Explain about Visualization tools used in Data science.
3. Explain the differences b/w Big data & Data science
4. Describe different types of statistical sampletests.
5. Explain the importance of EDA.
6. Describe different types of curve fitting model in statistics.

7. Explain about the process of Data science Activities
8. What is R Language, Explain its functions & Applications

10/02/25

Data Sciences

UNIT-3

Motivating customer Retension:-

Motivating customer Retention in an application for build the long term success, here some strategies to implement the retain users

1. Trailer content & offers: Use customer data to personali -ze the experience, offering relevant suggessions, content or discounts based on their behaviour & returns of the values
2. Push Notifications:- Send personalize remainders, updates or offers that make users feel & encourage them to return.
3. Recoards & baddebts:-Introduce points, recoards for completi -ng actions or achieving milestones within the app.
4. Leadership Qualities:-It is applicable to user Leadership qualfty, means frindly completions, motivating users to stay active to higher rank.
5. Exclusive Benefits:- Offer a loyalfty program where users can earn points, unlock exclusive content or receive spl discounts or promotions
6. Responsive support:- It used for offering & helpful customer to service the address of their encounter of the retailers.
7. Regular Updates:-It is used to continuously updating the app with new features & improvements to keep it fresh & user friendly.

8. Seasonal & timely content:- It is used to update with the current seasons offering users something new to look forward.

9. Educational content:- share content that helps, users get more out of the app for example tutorials, tips etc

10. Relevent offers:- It provides value based promotions such as time limited discounts or exclusively or early access to new features.

11. Easy to use:- Make sure your app is easy to use, quick to navigate & free of technical issues, smooth experience is key to retaining users

12. Social sharing:- It is used to encourage users to share their progress or experiences on social media & to build a sense of community.

13. User Feedback:- It is used to regularly to ask for feedback & involve users in decisions about app improvements & ownership & connections

14. Communication:- The users have smooth onboarding experiences that clearly shows how to use the app & while the should keep using it.

Ans. keep users informed with regular updates about new features or improvements & remind them of apps.

Filter models:-

Unit-IV

Data visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualization tools provides an acceable we provide to see and understand trends, outlayers, patterns in data.

Why Data visualization is important?

The importance of data visualization is simple, it helps people ~~is~~ s to see, ~~interupt~~ interact and better Understanding of data.

* Data visualization tools and technologies are essential to analyse massive amounts of information and, Data-driven decisions.
* while there are plenty of exoitic and interesting graph types, if you want to acurately show data patterns.
* Ideas for better data visualization

=> Choose the right chart type

* choosing the right chart type or defaulting to the most common type of data visualization could confuse users or lead to data misinterpretation
* If the same data set can be represented in many ways, depending on what users would like to see. Always start with a review of your dataset and User interview

=) User correct plotting directions based on positive and negative values

when using horizontal bars, plot negatives values on the left side and positive on the right side of a plot side. Donot plot negative and positive values

on the safe side of the base line.

Unit-V

Data Science and Ethical issues:

Data science has brought about significant advance-ments in various fields, from health care and finance to entertainment and education. However, as data science, relies heavily on data collection, processing and analysis, it raises several ethical issues. These ethical consens mainly revolve around the responsible handling, interpretation, and use of data, specially as its impact on individuals and society grows. some key ethical issues in data science inclu-de

privacy and data security:

* problem: with the large amounts of personal data being collected and analysed, consens around privacy and data security are at the forefornt this is especially critical in industries like health care, finance, and social media.
* Example: If sensitive health data (example patient records) is used without consent or is inadequately lead to data breches or misuse of personal information.

Ethical Consen: Ensuring that data is anonymised and securely stored, and that individual's concernt

Privacy

Privacy concerns how personal information is collected, stored, and used by individuals, companies, and governments. It refers to the ability of individuals to control their personal data and the conditions under which it is shared or made public. Key issues related to privacy include:

- **Data collection**: How much personal data is being collected and for what purposes?
- **Consent**: Are individuals aware of and have they agreed to how their data will be used?
- **Surveillance**: The balance between security and privacy, especially in terms of government surveillance programs.
- **Data breaches**: Protection against unauthorized access to personal data.

Security

Security refers to the protection of information systems and data from unauthorized access, use, disclosure, disruption, modification, or destruction. This includes:

- **Cybersecurity**: Measures to protect networks, devices, and data from cyberattacks, hacking, or malware.
- **Encryption**: Methods used to protect data by encoding it so only authorized parties can access it.
- **Authentication and access control**: Techniques used to verify the identity of users and restrict access to sensitive information.
- **Incident response**: How organizations respond to security breaches or data compro...

- **Cybersecurity**: Measures to protect networks, devices, and data from cyberattacks, hacking, or malware.
- **Encryption**: Methods used to protect data by encoding it so only authorized parties can access it.
- **Authentication and access control**: Techniques used to verify the identity of users and restrict access to sensitive information.
- **Incident response**: How organizations respond to security breaches or data compromises.

Ethics

Ethics in the context of privacy and security involves the moral implications of actions related to personal data, surveillance, and technology. This includes:

- **Responsibility**: Who is responsible for protecting personal data and ensuring ethical practices in handling it?
- **Transparency**: Are companies clear and honest about how they collect and use data?
- **Fairness**: Are all individuals treated equally and fairly in how their data is handled and protected?
- **Artificial Intelligence**: The ethical challenges of using AI to make decisions based on personal data, such as bias or discrimination in algorithms.

Each of these areas is interconnected, and finding the right balance between protecting personal privacy, ensuring security, and adhering to ethical standards is a complex challenge that continues to evolve as technology advances.

Filter Model

In data science, the term **filter model** generally refers to a technique or model that is used to select a subset of features, data points, or variables from a larger set based on certain criteria. The purpose is typically to reduce dimensionality or focus on the most important data to improve the efficiency and accuracy of a model. It can be applied to different areas such as feature selection, data preprocessing, or anomaly detection.

Here are a few common types of filter models in data science:

1. Filter Methods for Feature Selection

- **Feature selection** is the process of selecting a subset of relevant features for use in model construction. Filter methods are one way to achieve this.
- Filter methods evaluate each feature independently of the machine learning model, typically by using statistical measures such as correlation, mutual information, or chi-squared tests.
- These methods are called "filter" models because they "filter out" irrelevant features based on a certain criterion. They are usually simple, fast, and easy to interpret.
- Common statistical tests used for filter-based feature selection include:
    - **Correlation coefficient**: Measures the linear relationship between two variables.
    - **Chi-squared test**: Tests the independence between categorical variables.
    - **ANOVA (Analysis of Variance)**: Tests the difference in means between groups.
    - **Mutual Information**: Measures the dependence between two variables.

Example: If you are working with a dataset containing several features (e.g., age, income, education level), you might use a filter method to select only the features that are most correlated with the target variable (e.g., likelihood to buy a product).

Wrapper Models:-

A **wrapper model** in data science refers to a class of methods used for **feature selection** that evaluates subsets of features based on the performance of a specific machine learning model. Unlike **filter models**, which assess individual features based on statistical properties, wrapper models evaluate subsets of features by actually training and testing a model on those features, making them more computationally expensive but often more accurate in identifying the best features for a given task.

Key Characteristics of Wrapper Models:

1. **Feature Subset Evaluation**: Wrapper models evaluate different subsets of features and select the subset that leads to the best model performance. This is done by training the model on the subset and assessing its predictive power using cross-validation or another performance metric.
2. **Model Dependency**: The performance of the wrapper method is directly dependent on the machine learning model chosen for evaluation. For example, a wrapper method might use a decision tree, SVM, or logistic regression to assess how well a particular feature subset performs.
3. **Search Strategy**: Wrapper models use different search strategies to explore the feature space. The most common search strategies are:
    - **Exhaustive Search**: Tests all possible subsets of features. This method guarantees the optimal subset but can be computationally expensive, especially with many features.
    - **Heuristic Search (Greedy Search)**: Uses heuristics, such as forward or backward selection, to iteratively add or remove features based on model performance. These methods are computationally less expensive but might not always find the optimal solution.
    - **Random Search**: Selects random subsets of features and evaluates them. This method can be more efficient than exhaustive search, but it may not find the optimal solution.

Types of Wrapper Methods:

1. **Forward Selection**:
    - Starts with no features and adds one feature at a time that improves model performance the most. The process stops when adding another feature does not improve performance.
2. **Backward Elimination**:
    - Starts with all features and removes one feature at a time that reduces model performance the least. The process stops when removing a feature starts to degrade performance.
3. **Recursive Feature Elimination (RFE)**:
    - A popular wrapper method where features are recursively removed based on the performance of the model. In each iteration, features are ranked, and the least important feature is removed. The process continues until a predefined number of features is reached.

Example of How a Wrapper Model Works:

1. **Step 1**: Choose a model to train (e.g., logistic regression, decision tree, SVM).

2. **Step 2**: Generate a subset of features to evaluate. For example, start with an empty set of features (forward selection), or start with all features (backward elimination).
3. **Step 3**: Train the model using that subset of features.
4. **Step 4**: Evaluate model performance using a validation set or cross-validation (e.g., accuracy, F1-score, etc.).
5. **Step 5**: Adjust the feature subset by adding or removing features based on model performance.
6. **Step 6**: Repeat until the optimal subset of features is found.

Advantages of Wrapper Models:

1. **Better Performance**: Since wrapper methods are based on the actual performance of the model, they tend to find the most relevant features for the given learning algorithm, which often results in better performance compared to filter methods.
2. **Adaptability**: Wrapper methods can be used with any machine learning algorithm, so the feature selection process is tailored to the specific model used.
3. **Feature Interactions**: Wrapper models are capable of identifying interactions between features that might be overlooked by filter methods, as they assess subsets of features together.

Disadvantages of Wrapper Models:

1. **Computationally Expensive**: Evaluating multiple subsets of features by training and testing a model on each subset can be very computationally costly, especially for large datasets or complex models.
2. **Risk of Overfitting**: Because the subset selection is tied directly to the model performance, there's a risk of overfitting to the training data. Cross-validation is typically used to mitigate this risk, but it's still a potential concern.
3. **Model Dependency**: The performance of the wrapper method depends heavily on the chosen model. If a model is poorly selected for the task, the feature selection process may not yield good results.

Example Use Case of Wrapper Models:

- **Predicting Customer Churn**: Suppose you want to predict whether a customer will churn (leave your service) based on features like age, income, usage patterns, etc. A wrapper model might use a logistic regression or decision tree to iteratively select the best features that contribute the most to predicting customer churn. It would evaluate different subsets of features (e.g., age + income, usage patterns + income, etc.) and choose the subset that gives the highest accuracy or F1-score for the chosen model.

Conclusion:

Wrapper models are a powerful feature selection method that can help improve the performance of machine learning models by selecting the most relevant features based on model-specific performance. While they tend to provide better feature subsets than filter methods, they come at the cost of higher computational resources and longer training times. They are ideal when computational resources are available, and the goal is to optimize the performance of a specific machine learning algorithm.

Feature Generation Algorithm

Feature generation (also known as **feature engineering**) in machine learning involves creating new features from the existing ones to improve the model's predictive power. It is a crucial step in the data preprocessing pipeline, as the quality and variety of features directly influence the performance of machine learning models.

Feature generation techniques can help create new insights from raw data, combine different variables, or extract hidden patterns to make the model more effective. These techniques can be applied to numerical, categorical, and text data.

Here's a breakdown of common **feature generation algorithms** with examples:

1. Polynomial Features (Interaction Features)

- **Description**: This involves creating new features by taking combinations of existing features. For instance, if you have two features x1x_1x1 and x2x_2x2, polynomial features can generate new features like x12x_1^2x12, x22x_2^2x22, and x1·x2x_1 \cdot x_2x1·x2.
- **Use Case**: Useful in linear models to capture interactions between variables that may not be linearly related.
- **Example**:
  - Original features: `age`, `income`
  - Generated features: `age^2`, `income^2`, `age * income`

Example in Python using `PolynomialFeatures` from `sklearn`:

```
python
Copy
from sklearn.preprocessing import PolynomialFeatures

# Original features: age and income
X = [[25, 50000], [30, 60000], [35, 55000]]

poly = PolynomialFeatures(degree=2)
X_poly = poly.fit_transform(X)
print(X_poly)
```

This will generate the following features: 1,age,income,age2,income2,age·income1, age, income, age^2, income^2, age \cdot income1,age,income,age2,income2,age·income.

2. Binning (Discretization)

- **Description**: Converts continuous variables into categorical ones by grouping them into bins or intervals. This can help capture nonlinear relationships.
- **Use Case**: Often used for numerical features, especially when they have outliers or are not normally distributed.
- **Example**:
  - Original feature: `age` (continuous)
  - Generated feature: `age_group` (discrete), where age is categorized as: "Young", "Middle-aged", "Old"

Example in Python using `KBinsDiscretizer`:

```
python
Copy
from sklearn.preprocessing import KBinsDiscretizer

X = [[25], [30], [35], [40], [45]]
binner = KBinsDiscretizer(n_bins=3, encode='ordinal', strategy='uniform')
X_binned = binner.fit_transform(X)
print(X_binned)
```

This would divide the `age` variable into 3 bins based on the distribution and encode them as discrete values.

3. Date/Time Features (Temporal Features)

- **Description**: Extracting features from date/time data such as the day of the week, month, year, or time of day.
- **Use Case**: Important in time series forecasting, and also useful in any predictive model involving temporal data.
- **Example**:
  - Original feature: `purchase_date` (timestamp)
  - Generated features: `day_of_week`, `month`, `hour_of_day`, `is_weekend`

Example in Python:

```
python
Copy
import pandas as pd

# Original date feature
df = pd.DataFrame({'purchase_date': ['2025-02-01 14:30', '2025-02-02
09:15', '2025-02-03 20:45']})
df['purchase_date'] = pd.to_datetime(df['purchase_date'])

# Extract temporal features
df['day_of_week'] = df['purchase_date'].dt.dayofweek
df['month'] = df['purchase_date'].dt.month
df['hour_of_day'] = df['purchase_date'].dt.hour
df['is_weekend'] = df['purchase_date'].dt.weekday >= 5

print(df)
```

This will add columns like `day_of_week`, `month`, `hour_of_day`, and `is_weekend` to the data.

4. Text Feature Generation

- **Description**: Text data can be transformed into numerical features using various techniques like **TF-IDF (Term Frequency-Inverse Document Frequency)**, **Word2Vec**, or **Bag of Words**.
- **Use Case**: Essential in Natural Language Processing (NLP) tasks like sentiment analysis, document classification, or text summarization.
- **Example**:
  - Original feature: `review_text` (text)
  - Generated features: `word_frequency`, `TF-IDF score`, `sentiment_score`

Example using `TfidfVectorizer` in Python:

```
python
Copy
from sklearn.feature_extraction.text import TfidfVectorizer

reviews = ["Great product", "Not bad", "Could be better", "Loved it"]
vectorizer = TfidfVectorizer()
X_tfidf = vectorizer.fit_transform(reviews)

print(X_tfidf.toarray())
```

This will create a matrix where each row represents a document, and each column corresponds to the TF-IDF score of a specific word across the documents.

5. Aggregating Features (Statistical Features)

- **Description**: Creating features by aggregating values over a period or group, such as mean, median, standard deviation, or max.
- **Use Case**: Often used in time series or grouped data, where statistics from previous time steps or groupings provide valuable information.
- **Example**:
  - Original features: `sales` over time
  - Generated features: `mean_sales_last_7_days`, `max_sales_last_month`

Example using `groupby` and `agg` in Python:

```
python
Copy
import pandas as pd

# Sample data
data = {'store_id': [1, 1, 1, 2, 2, 2],
        'sales': [100, 150, 200, 80, 120, 180],
        'date': pd.to_datetime(['2025-01-01', '2025-01-02', '2025-01-03',
'2025-01-01', '2025-01-02', '2025-01-03'])}
df = pd.DataFrame(data)

# Aggregating sales for each store
df['mean_sales'] = df.groupby('store_id')['sales'].transform('mean')
df['max_sales'] = df.groupby('store_id')['sales'].transform('max')

print(df)
```

This generates statistical features like `mean_sales` and `max_sales` based on the grouping of `store_id`.

6. Domain-Specific Feature Generation

- **Description**: This involves creating new features based on domain knowledge. For instance, in a financial dataset, you could generate features like the **debt-to-income ratio**, which is a combination of two existing features.
- **Use Case**: Particularly useful when you have a strong understanding of the domain you're working in.

- **Example**:
  - Original features: `debt, income`
  - Generated feature: `debt_to_income_ratio`

Example:

```
python
Copy
df['debt_to_income_ratio'] = df['debt'] / df['income']
```

Conclusion:

Feature generation plays a significant role in improving the performance of machine learning models by creating new, informative features from raw data. The examples above cover a variety of approaches for generating features from numerical, temporal, and text data. Each technique should be chosen based on the problem you are trying to solve and the type of data you are working with. Effective feature generation can make a considerable difference in the performance of predictive models, especially when combined with good feature selection.