

BUSINESS STATISTICS



Types of Data Presentation



Pictorial

Graphical

Bar Diagram

Pie Diagram

Data Array

Pictograph

Histogram

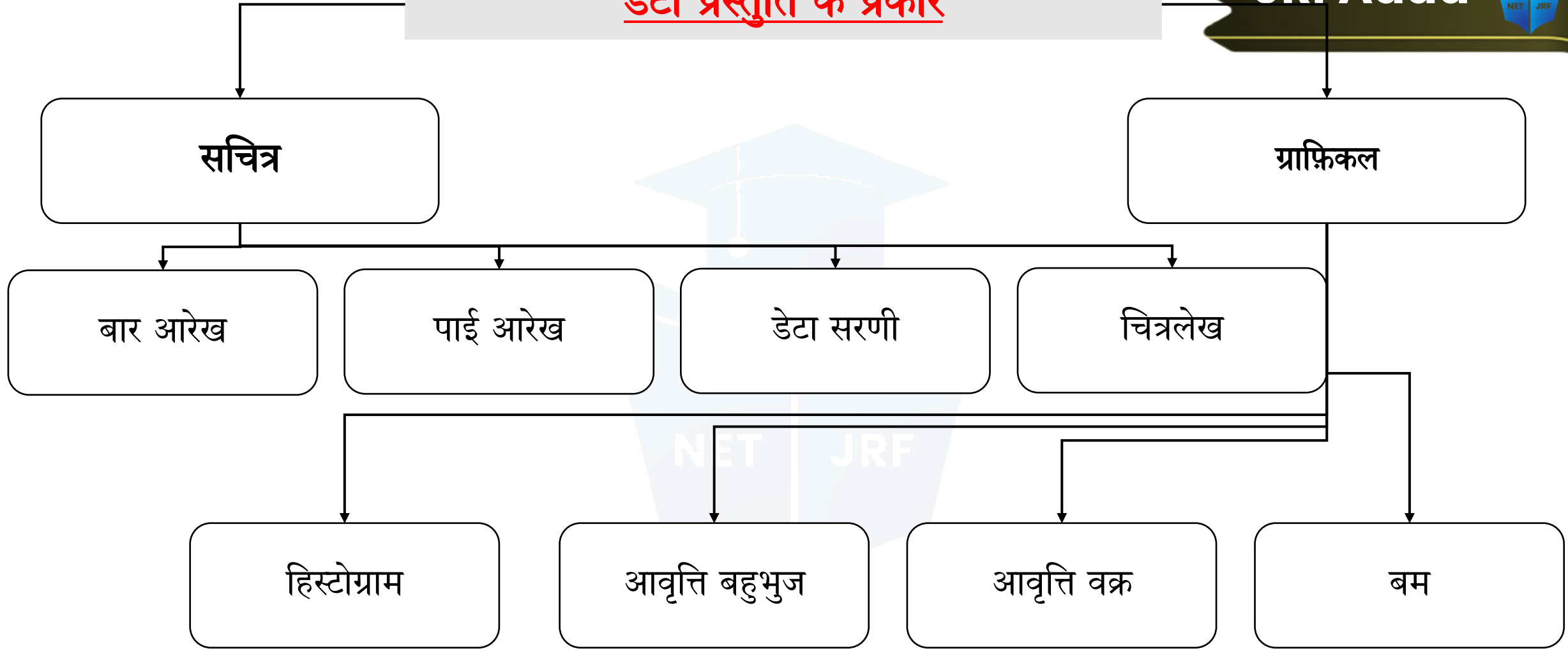
Frequency
Polygon

Frequency
Curve

Ogive

डेटा प्रस्तुति के प्रकार

JRFAdda

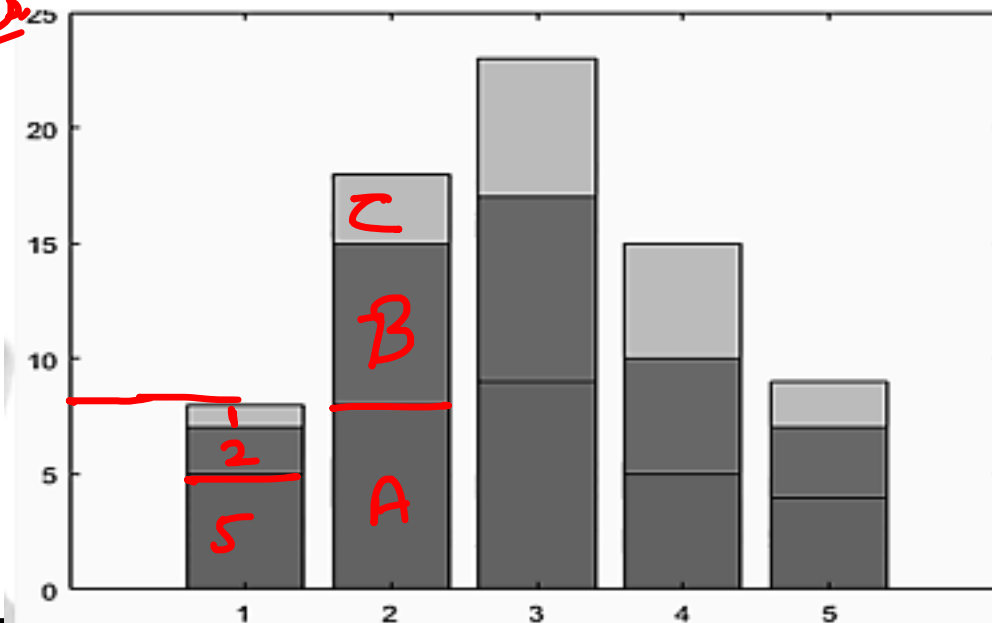


A **Bar Chart** or **Bar Graph** is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

बार चार्ट या बार ग्राफ़ एक चार्ट या ग्राफ़ है जो श्रेणीबद्ध डेटा को आयताकार पट्टियों के साथ प्रस्तुत करता है जिनकी ऊँचाई या लंबाई उनके द्वारा दर्शाए गए मानों के समानुपाती होती है। सलाखों को लंबवत या क्षैतिज रूप से प्लॉट किया जा सकता है। वर्टिकल बार चार्ट को कभी-कभी कॉलम चार्ट भी कहा जाता है।

Handwritten notes in red ink:

- 100%
- M
- 2 : 3
- 60%
- lakh



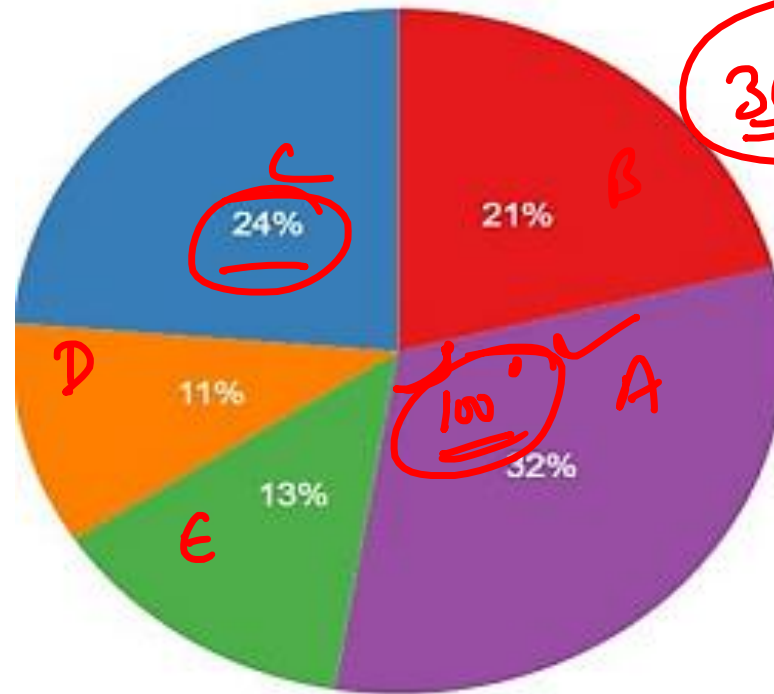
Handwritten notes in red ink:

- 2021
- 2022
- 2023

A Pie Chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice, is proportional to the quantity it represents.

पाई चार्ट एक गोलाकार सांख्यिकीय ग्राफिक है, जिसे संख्यात्मक अनुपात को दर्शाने के लिए स्लाइस में विभाजित किया गया है। पाई चार्ट में, प्रत्येक स्लाइस की चाप लंबाई, उसके द्वारा दर्शाई गई मात्रा के समानुपाती होती है।

12000



360°

$$\frac{100}{360} \times 100$$

A/c
of
Stats
Income

Table
Meth
MEM-BH

डेटा सारणी

DATA ARRAY

An Array is a systematic arrangement of objects, usually in rows and columns Data Array- Observations that are systematically arranged.

सारणी वस्तुओं की एक व्यवस्थित व्यवस्था है, आमतौर पर पंक्तियों और स्तंभों में डेटा सारणी- अवलोकन जो व्यवस्थित रूप से व्यवस्थित होते हैं।

Arrays in Data Structure

One-Dimensional Array




Index	1	2	3	4	5
Value	11	25	89	245	65

	Index	1	2	3	
					55
					96
				38	75
				65	
				90	
Index	1	2	3		
1	100	102	104		
2	9	25	45		
3	75	2	83		

Multi Dimensional Array

A Pictogram, also called a pictograms, pictograph, or simply picto, and in computer usage an icon, is a graphic symbol that conveys its meaning through its pictorial resemblance to a physical object.

एक पिक्टोग्राम, जिसे पिक्टोग्राम, पिक्टोग्राफ, या बस पिक्टो भी कहा जाता है, और कंप्यूटर के उपयोग में एक आइकन, एक ग्राफिक प्रतीक है जो किसी भौतिक वस्तु के चित्रात्मक समानता के माध्यम से अपना अर्थ बताता है।

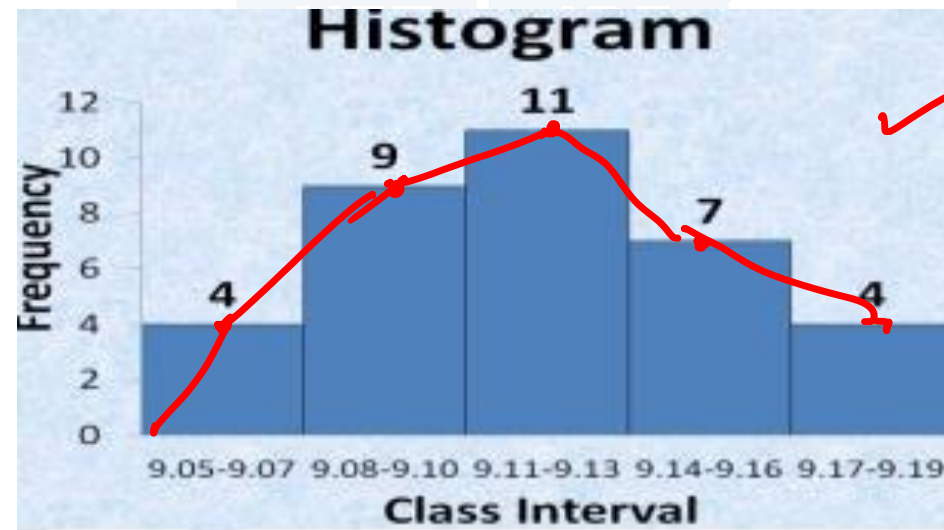
Green Apples	
Red Apples	
Yellow Apples	

QUI

Icon

A **Histogram** is a bar graph-like representation of data that buckets a range of outcomes into columns along the x-axis. The y-axis represents the number count or percentage of occurrences in the data for each column and can be used to visualize data distributions.

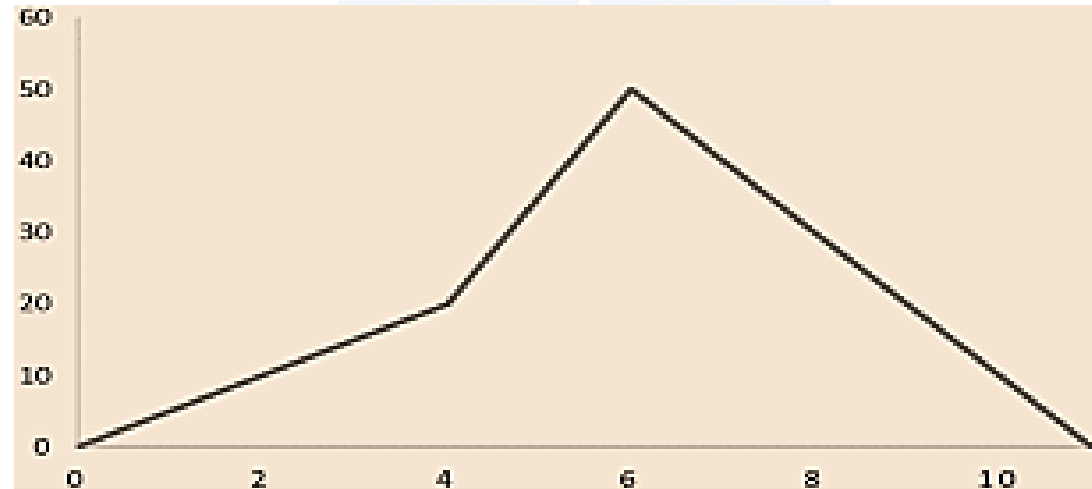
हिस्टोग्राम डेटा का एक बार ग्राफ जैसा प्रतिनिधित्व है जो एक्स-अक्ष के साथ कॉलम में परिणामों की एक श्रृंखला को बकेट करता है। y-अक्ष प्रत्येक कॉलम के लिए डेटा में संख्या गणना या घटनाओं के प्रतिशत का प्रतिनिधित्व करता है और इसका उपयोग डेटा वितरण को देखने के लिए किया जा सकता है।



A **Frequency Polygon** is a visual representation of a distribution. The visualization tool is used to understand the shape of a distribution.

Essentially, the frequency polygon indicates the number of occurrences for each distinct class in the dataset ✓

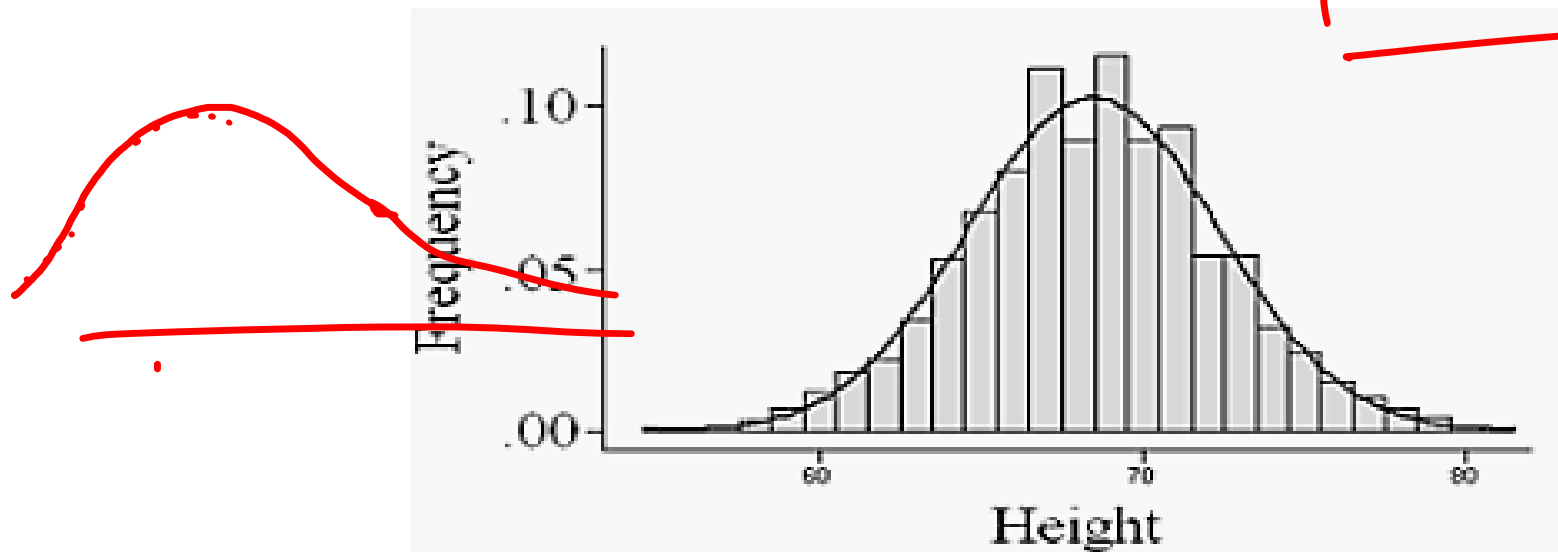
बारंबारता बहुभुज एक वितरण का एक दृश्य प्रतिनिधित्व है। विज़ुअलाइज़ेशन टूल का उपयोग वितरण के आकार को समझने के लिए किया जाता है। अनिवार्य रूप से, आवृत्ति बहुभुज डेटासेट में प्रत्येक विशिष्ट वर्ग के लिए घटनाओं की संख्या को इंगित करता है



A.H!

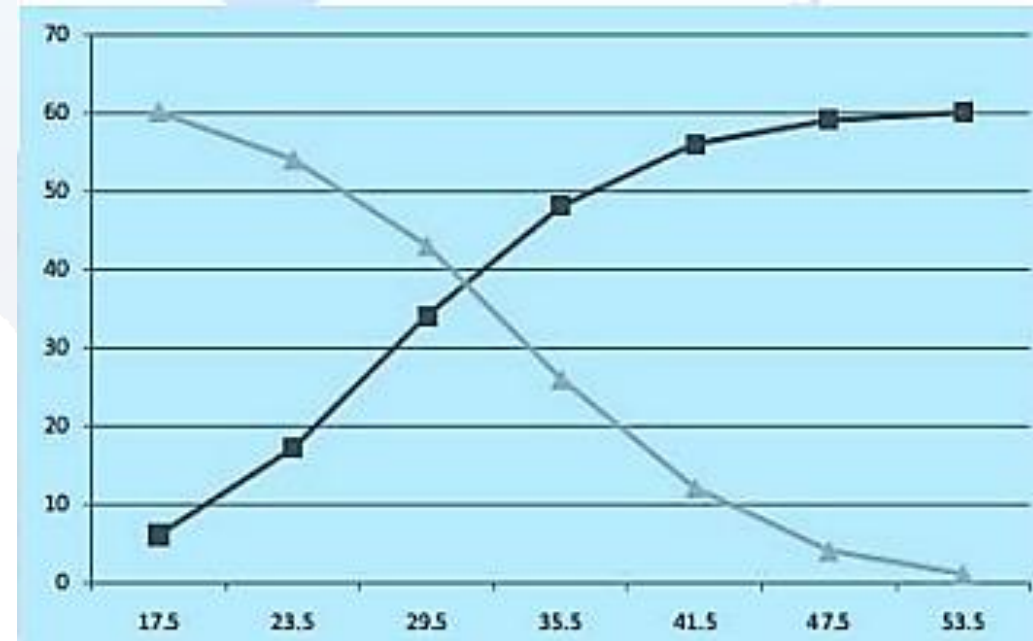
A **Frequency Curve** is a smooth curve which corresponds to the limiting case of a histogram computed for a frequency distribution of a continuous distribution as the number of data points becomes very large.

फ्रीक्वेंसी वक्र एक चिकना वक्र है जो निरंतर वितरण के आवृत्ति वितरण के लिए गणना किए गए हिस्टोग्राम के सीमित मामले से मेल खाता है क्योंकि डेटा बिंदुओं की संख्या बहुत बड़ी हो जाती है।



An **Ogive** sometimes called a cumulative frequency polygon, is a type of frequency polygon that shows cumulative frequencies. In other words, the cumulative percents are added on the graph from left to right. An ogive graph plots cumulative frequency on the y-axis and class boundaries along the x-axis.

एक तोरण जिसे कभी-कभी संचयी आवृत्ति बहुभुज भी कहा जाता है, एक प्रकार का आवृत्ति बहुभुज है जो संचयी आवृत्तियों को दर्शाता है। दूसरे शब्दों में, संचयी पूर्वसर्ग ग्राफ़ पर बाएँ से दाएँ जोड़े जाते हैं। एक तोरण ग्राफ़ y-अक्ष पर संचयी आवृत्ति और x-अक्ष के साथ वर्ग सीमाओं को आलेखित करता है।





Sturges Rule

- Sturges' Rule uses the following formula to determine the optimal number of bins (rectangular bars) to use in a histogram: **Optimal Bins $\Rightarrow K = [1 + 3.322 \log N]$**
- Where, k = the number of bins, n = the number of observations in the data set.
- Symbols that mean "ceiling" - i.e. round the answer up to the nearest integer.
- It turns out that the number of bins used in a histogram can have a huge impact on how we interpret the data.



स्टर्जेस नियम

- हिस्टोग्राम में उपयोग करने के लिए डिब्बे (आयताकार बार) की इष्टतम संख्या निर्धारित करने के लिए स्टर्जेस नियम निम्नलिखित सूत्र का उपयोग करता है: इष्टतम डिब्बे $\Rightarrow k = [1 + 3.322 \log n]$
- जहां, k = डिब्बे की संख्या, n = डेटा सेट में अवलोकनों की संख्या।
- [1: ऐसे प्रतीक जिनका अर्थ है "छत" - यानी उत्तर को निकटतम पूर्णांक तक पूर्णांकित करें।
- यह पता चला है कि हिस्टोग्राम में उपयोग किए गए डिब्बे की संख्या इस बात पर भारी प्रभाव डाल सकती है कि हम डेटा की व्याख्या कैसे करते हैं।



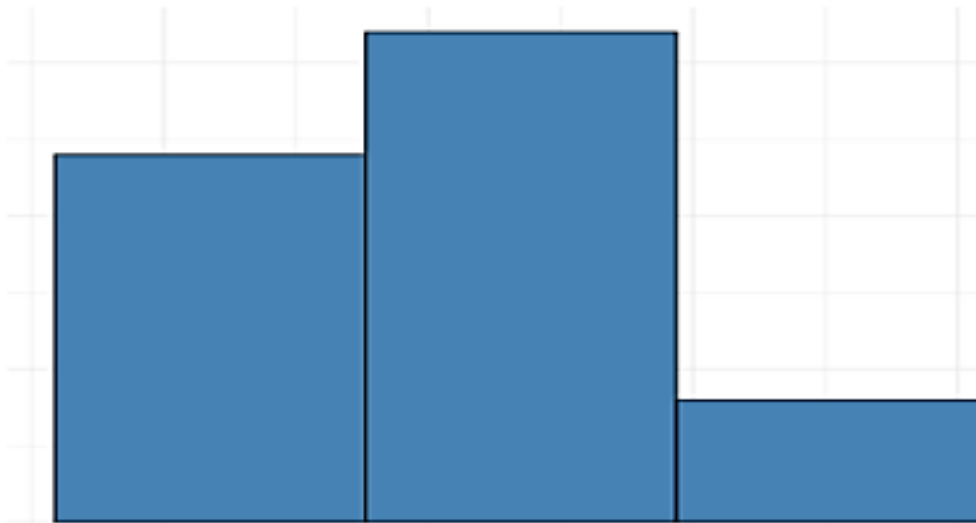
If we use too **few bins**, the **true** underlying pattern in the data can be **hidden**:

यदि हम बहुत कम डिब्बे का उपयोग करते हैं, तो डेटा में वास्तविक अंतर्निहित पैटर्न छिपाया जा सकता है:

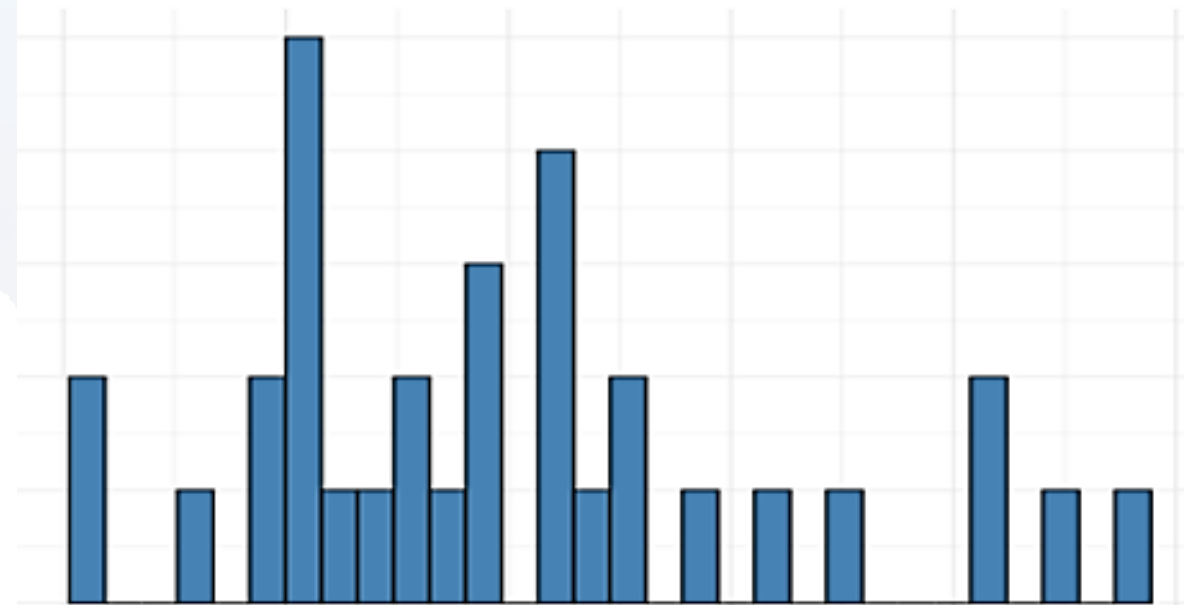
And if we use **too many bins**, we may just be visualizing the **noise in a dataset**:

और यदि हम बहुत अधिक डिब्बे का उपयोग करते हैं, तो हम केवल डेटासेट में शोर की कल्पना कर सकते हैं:

Total Bins = 3



Total Bins = 30



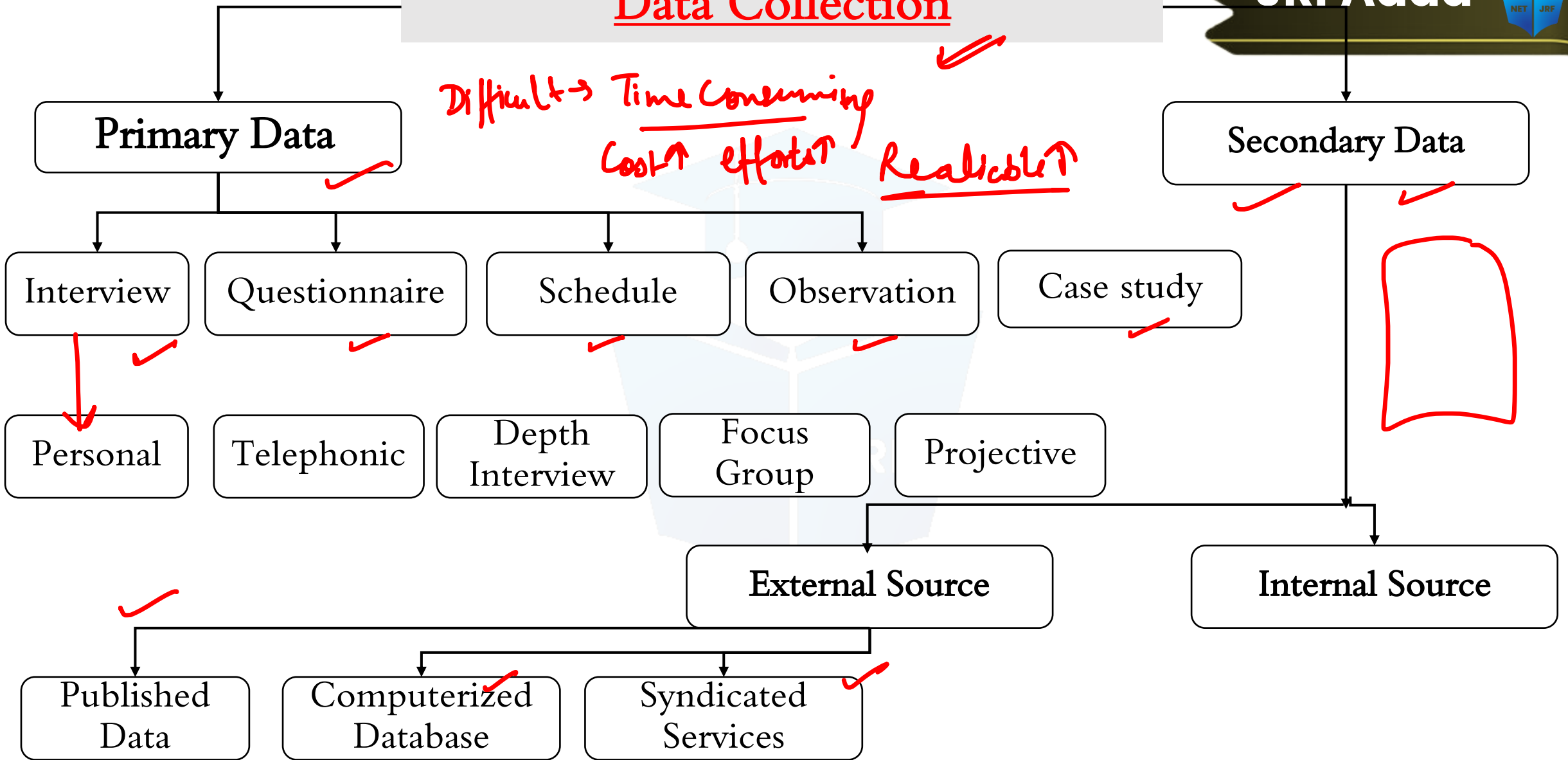


Open-Ended	Close-Ended	Exclusive	Inclusive
<p>This gives the respondents the ability to respond in their own words. (Subjective)</p>	<p>This allows the respondents to choose one of the given alternatives. (Objective)</p>	<p>In this we include lower limit but exclude upper limit.</p>	<p>Value of upper limit of a class is not included in the lower limit of the next interval.</p>
<p><u>For Example-</u></p> <p>0 > 5 5 > 10 10 > 15 15 > 20 20 > 25 25 > 30</p> <p>(Eg: In this series we can take any number which is less than 5 in first interval and so on)</p>	<p><u>For Example-</u></p> <p>0-5 5-10 10-15 15-20 20-25 25-30</p> <p>(Eg: In this series fixed interval is given we can't go beyond that limit.)</p>	<p><u>For Example-</u></p> <p>0-5 5-10 10-15 15-20 20-25 25-30</p> <p>(Eg: In this series we include lower but exclude upper limit)</p>	<p><u>For Example-</u></p> <p>0-5 6-10 11-15 16-20 ✓ 21-25 26-30</p> <p>(Eg- 5 (upper limit) of 0 - 5 series is not included in 6 (lower limit) of 6-10 series)</p>

खुला	क्लोज-एंडेड	निवारक	सहित
यह उत्तरदाताओं को अपने शब्दों में जवाब देने की क्षमता देता है। (व्यक्तिपरक)	यह उत्तरदाताओं को दिए गए विकल्पों में से एक को चुनने की अनुमति देता है। (उद्देश्य)	इसमें हम निचली सीमा को शामिल करते हैं लेकिन ऊपरी सीमा को बाहर करते हैं।	किसी कक्षा की ऊपरी सीमा का मान अगले अंतराल की निचली सीमा में शामिल नहीं है।
<u>उदाहरण के लिए-</u> $0 > 5$ $5 > 10$ $10 > 15$ $15 > 20$ $20 > 25$ $25 > 30$ (उदाहरण के लिए: इस श्रृंखला में हम किसी भी संख्या को ले सकते हैं जो पहले अंतराल में 5 से कम है और इसी तरह)	<u>उदाहरण के लिए-</u> $0-5$ $5-10$ $10-15$ $15-20$ $20-25$ $25-30$ (जैसे: इस श्रृंखला में निश्चित अंतराल दिया गया है हम उस सीमा से आगे नहीं जा सकते हैं।)	<u>उदाहरण के लिए-</u> $0-5$ $5-10$ $10-15$ $15-20$ $20-25$ $25-30$ (जैसे: इस श्रृंखला में हम निम्न लेकिन ऊपरी सीमा को शामिल नहीं करते हैं)	<u>उदाहरण के लिए-</u> $0-5$ $6-10$ $11-15$ $16-20$ $21-25$ $26-30$ (उदाहरण के लिए- 0 - 5 श्रृंखला की 5 (ऊपरी सीमा) 6-10 श्रृंखला के 6 (निचली सीमा) में शामिल नहीं है)

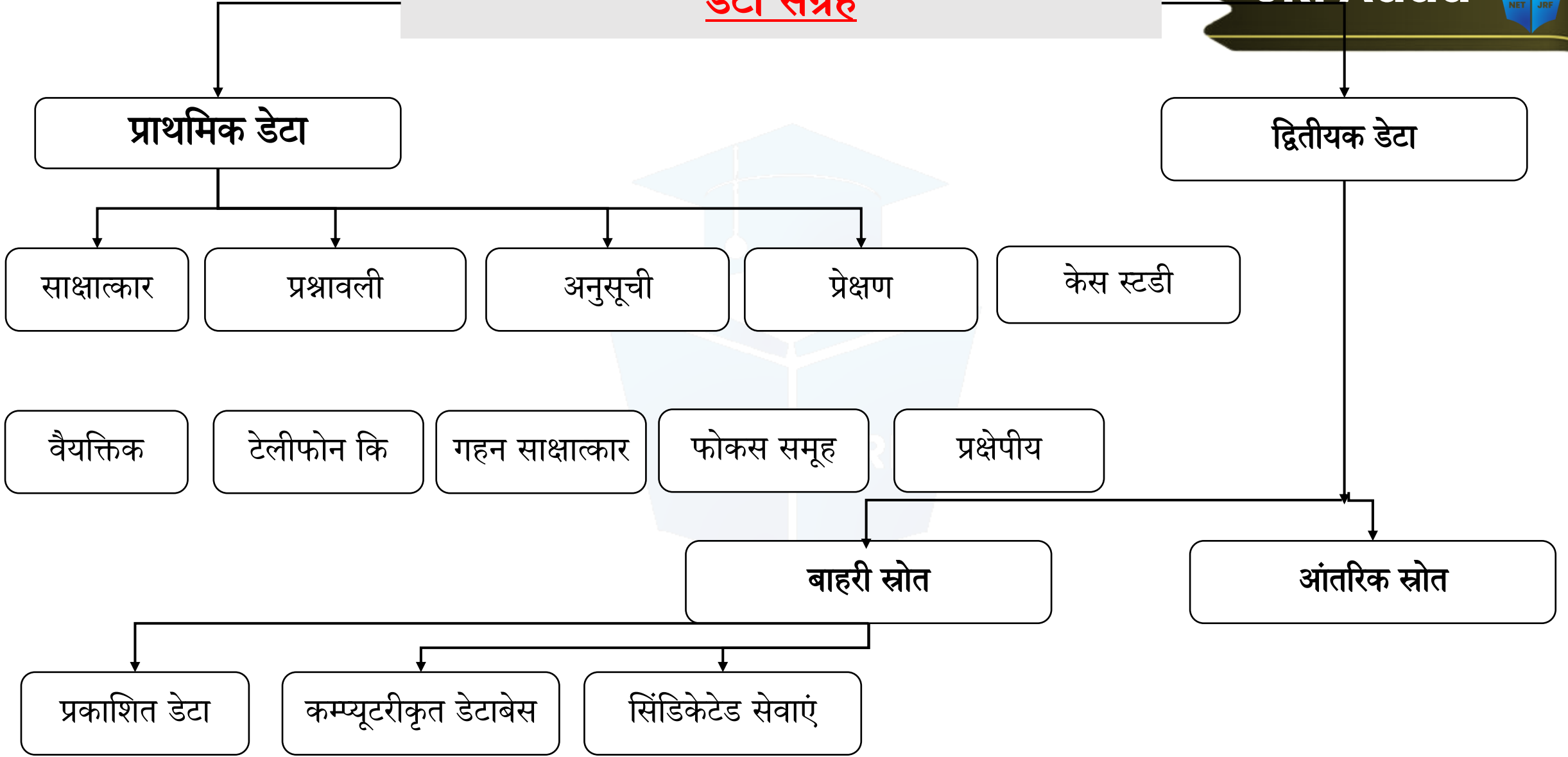
Data Collection

Difficult → Time Consuming
Cost ↑ Effort ↑
Realistic ↑



डेटा संग्रह

JRFAdda





PRIMARY DATA

Primary data is a type of data that is collected by researchers directly from main sources through interviews, surveys, experiments, etc. Primary data are usually collected from the source.

प्राथमिक डेटा

प्राथमिक डेटा एक प्रकार का डेटा है जिसे शोधकर्ताओं द्वारा साक्षात्कार, सर्वेक्षण, प्रयोग आदि के माध्यम से सीधे मुख्य स्रोतों से एकत्र किया जाता है। प्राथमिक डेटा आमतौर पर स्रोत से एकत्र किया जाता है।

Advantages Of Primary Data:

- Targeted Issues are addressed ✓
- Data interpretation is better ✓
- Efficient Spending for Information ✓
- Decency of Data ✓
- Proprietary Issues ✓
- Addresses Specific Research Issues ✓
- Greater Control ✓

प्राथमिक डेटा के लाभ:

- लक्षित मुद्दों का समाधान किया जाता है
- डेटा इंटरप्रिटेशन बेहतर है
- सूचना के लिए कुशल व्यय
- डेटा की शालीनता
- मालिकाना मुद्दे
- विशिष्ट अनुसंधान मुद्दों को संबोधित करता है
- अधिक नियंत्रण

Disadvantages Of Primary Data:

- There are numerous hassles involved in the collection of primary data like taking a decision such as how, when, what and why to collect.
- The cost involved in the collection of primary data is very high
- More number of resources are required.
- Inaccurate feedbacks

प्राथमिक डेटा के नुकसान:

- प्राथमिक डेटा के संग्रह में कई झंझटें शामिल हैं जैसे कि कैसे, कब, क्या और क्यों एकत्र करना है जैसे निर्णय लेना।
- प्राथमिक डेटा के संग्रह में शामिल लागत बहुत अधिक है
- अधिक संख्या में संसाधनों की आवश्यकता है.
- ग़लत प्रतिक्रियाएँ

SECONDARY DATA

Secondary data is the data that has already been collected through primary sources and made readily available for researchers to use for their own research. It is a type of data that has already been collected in the past

सहायक डेटा

द्वितीयक डेटा वह डेटा है जो पहले से ही प्राथमिक स्रोतों के माध्यम से एकत्र किया गया है और शोधकर्ताओं को अपने स्वयं के शोध के लिए उपयोग करने के लिए आसानी से उपलब्ध कराया गया है। यह एक प्रकार का डेटा है जिसे पहले ही एकत्र किया जा चुका है

Advantages Of Secondary Data:

- May have fewer resource requirements
- Unobtrusive
- Longitudinal studies may be feasible
- Can provide comparative and contextual data Can result in unforeseen discoveries and insights
- Permanence of data

द्वितीयक डेटा के लाभ:

- कम संसाधन आवश्यकताएँ हो सकती हैं
- विनीत
- अनुदैर्घ्य अध्ययन संभव हो सकता है
- तुलनात्मक और प्रासंगिक डेटा प्रदान कर सकता है जिसके परिणामस्वरूप अप्रत्याशित खोजें और अंतर्दृष्टि प्राप्त हो सकती हैं
- डेटा का स्थायित्व

2019 2020

Comparison

Disadvantages Of Secondary Data :

- May be collected for a purpose that does not match your need
- Access may be difficult or costly
- Aggregations and definitions may be unsuitable
- No real control over data quality
- Initial purpose may affect how data are presented

द्वितीयक डेटा के नुकसान:

- किसी ऐसे उद्देश्य के लिए एकत्र किया जा सकता है जो आपकी आवश्यकता से मेल नहीं खाता
- पहुंच कठिन या महंगी हो सकती है
- एकीकरण और परिभाषाएं अनुपयुक्त हो सकती हैं
- डेटा गुणवत्ता पर कोई वास्तविक नियंत्रण नहीं
- प्रारंभिक उद्देश्य डेटा प्रस्तुत करने के तरीके को प्रभावित कर सकता है



K K South

Primary V/S Secondary Data

PRIMARY DATA	SECONDARY DATA
Data collected by researcher himself	Data already collected by any other person
Original or unique information	Not original or unique information
Surveys, observations, experiments are example	Internal record, government published data etc.
More time consuming	Less time consuming
More Expensive and more reliable	Less Expensive and less reliable
Needs team of trained investigators	No need of team of investigators
Most suitable to achieve objective	May or may not be suitable
Collected when secondary data is insufficient	Collected before primary data
Possibility of bias exist	X Somewhat safe from bias
Precaution to use not necessary	Precaution to use quite necessary
Doesn't need adjustments	Needs adjustment to suit actual aim



प्राथमिक डेटा	द्वितीयक डेटा
शोधकर्ता द्वारा स्वयं एकत्र किया गया डेटा	किसी अन्य व्यक्ति द्वारा पहले से एकत्र किया गया डेटा
मूल या अद्वितीय जानकारी	मूल या अद्वितीय जानकारी नहीं
सर्वेक्षण, अवलोकन, प्रयोग उदाहरण हैं।	आंतरिक रिकॉर्ड, सरकार द्वारा प्रकाशित डेटा आदि।
अधिक समय लेने वाला	कम समय लेने वाला
अधिक महंगा और अधिक विश्वसनीय	कम महंगा और कम विश्वसनीय
प्रशिक्षित जांचकर्ताओं की टीम की जरूरत	जांचकर्ताओं की टीम की जरूरत नहीं
उद्देश्य को प्राप्त करने के लिए सबसे उपयुक्त	उपयुक्त हो भी सकता है और नहीं भी।
द्वितीयक डेटा अपर्याप्त होने पर एकत्र किया जाता है	प्राथमिक डेटा से पहले एकत्र किया गया
पूर्वाग्रह की संभावना मौजूद है	पूर्वाग्रह से कुछ हद तक सुरक्षित
उपयोग करने के लिए सावधानी आवश्यक नहीं है	काफी आवश्यक उपयोग करने के लिए सावधानी
समायोजन की आवश्यकता नहीं है	वास्तविक लक्ष्य के अनुरूप समायोजन की आवश्यकता है



INTERVIEW

Personal Interview: It involves the gathering of data through direct verbal interaction between interviewer and interviewee.

Telephone Interview: the information is collected from the respondent by asking him questions on the phone.

साक्षात्कार

व्यक्तिगत साक्षात्कार: इसमें साक्षात्कारकर्ता और साक्षात्कारकर्ता के बीच सीधे मौखिक बातचीत के माध्यम से डेटा एकत्र करना शामिल है।

टेलीफोन साक्षात्कार: उत्तरदाता से फोन पर प्रश्न पूछकर जानकारी एकत्र की जाती है।

Interviewer

Questions

Depth Interview: is a qualitative research technique that conducts intensive individual interview with a small number of respondents to explore their perspectives on a particular idea, program, situation etc.

गहन साक्षात्कार: एक गुणात्मक अनुसंधान तकनीक है जो किसी विशेष विचार, कार्यक्रम, स्थिति आदि पर उनके दृष्टिकोण का पता लगाने के लिए कम संख्या में उत्तरदाताओं के साथ गहन व्यक्तिगत साक्षात्कार आयोजित करती है।

Focus Group Interview: involves gathering people from similar background or experiences together to discuss a specific topic of interest.

फोकस समूह साक्षात्कार: इसमें रुचि के किसी विशिष्ट विषय पर चर्चा करने के लिए समान पृष्ठभूमि या अनुभव वाले लोगों को एक साथ इकट्ठा करना शामिल है।

NEP 2020

Housewife

Shopkeeper

NEP 2020

BackJolly

QUESTIONNAIRE	SCHEDULE
Sent through mail	Filled out by the researcher
Relatively economical	Relatively more expensive
Chance of non-response is high	Chance of non-response is low
Respondent is unknown	Respondent is known
No Personal contact	Personal contact is possible
Useful only when respondent is literate	Possible even respondents are illiterate

प्रश्नावली	अनुसूची
मेल के माध्यम से भेजा	शोधकर्ता द्वारा भरा गया
अपेक्षाकृत किफायती	अपेक्षाकृत अधिक महंगा
गैर-प्रतिक्रिया की संभावना अधिक है	गैर-प्रतिक्रिया की संभावना कम है
उत्तरदाता अज्ञात है	उत्तरदाता ज्ञात है
कोई व्यक्तिगत संपर्क नहीं	व्यक्तिगत संपर्क संभव है।
केवल तभी उपयोगी होता है जब उत्तरदाता साक्षर हो	यहां तक कि उत्तरदाता भी अनपढ़ हैं।

Quantitative

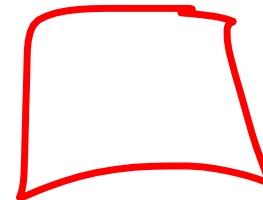
Observations: refers to monitoring & recording of behavioural & non-behavioural activities and conditions in a systematic manner to obtain the needed information about the event of interest.

Case Study: can be defined as an intensive study about a person, a group of people or a unit, which is aimed to generalize over several units'.

अवलोकन: रुचि की घटना के बारे में आवश्यक जानकारी प्राप्त करने के लिए व्यवस्थित तरीके से व्यवहारिक और गैर-व्यवहारिक गतिविधियों और स्थितियों की निगरानी और रिकॉर्डिंग को संदर्भित करता है।

केस स्टडी: इसे किसी व्यक्ति, लोगों के समूह या एक इकाई के बारे में गहन अध्ययन के रूप में परिभाषित किया जा सकता है, जिसका उद्देश्य कई इकाइयों का सामान्यीकरण करना है।

generative



INTERNAL SOURCE OF SECONDARY DATA द्वितीयक डेटा का आंतरिक स्रोत

Internal Secondary Data: consist of information gathered else- where within your firm. The major categories include ✓

- (1) sales reports, ✓
- (2) customer databases etc. ✓

आंतरिक माध्यमिक डेटा: इसमें आपकी फर्म के भीतर कहीं और एकत्रित की गई जानकारी शामिल होती है। प्रमुख श्रेणियाँ शामिल हैं

- (1) बिक्री रिपोर्ट,
- (2) ग्राहक डेटाबेस आदि।

EXTERNAL SOURCE OF SECONDARY DATA द्वितीयक डेटा का बाहरी स्रोत

✓ Published Data: sources of information prepared for public distribution and normally found in libraries or a variety of other entities, such as trade organization.

प्रकाशित डेटा: सार्वजनिक वितरण के लिए तैयार की गई जानकारी के स्रोत और आम तौर पर पुस्तकालयों या व्यापार संगठन जैसी कई अन्य संस्थाओं में पाए जाते हैं।

Index

RBI

Reports

external

Computerized Data: is a collection of numeric data and/or textual information that is made available in computer-readable form for electronic distribution.

Syndicated Services: *Info collect* are companies that collect and sell common data of known commercial value designed to serve a number of clients.

कम्प्यूटरीकृत डेटा: संख्यात्मक डेटा और/या पाठ्य सूचना का एक संग्रह है जिसे इलेक्ट्रॉनिक वितरण के लिए कंप्यूटर-पठनीय रूप में उपलब्ध कराया जाता है।

सिंडिकेटेड सेवाएँ: वे कंपनियाँ हैं जो कई ग्राहकों की सेवा के लिए डिज़ाइन किए गए ज्ञात वाणिज्यिक मूल्य का सामान्य डेटा एकत्र करती हैं और बेचती हैं।

CENSUS <u>10</u>	SURVEY
Census collects information about <u>every member</u> of the population	Survey collects information from a <u>sample</u> of the population
Census is more <u>detailed and accurate</u>	Survey is <u>not</u> accurate or reliable as a Census
Census takes a <u>long time</u> to complete	Surveys can be <u>done in a shorter</u> period of time compared to a census
Census is generally <u>conducted by the government</u>	Surveys can be <u>conducted by anyone</u>
Census are <u>not conducted frequently</u> (after every 10 years)	Surveys can be <u>conducted more frequently.</u>

जनगणना	सर्वेक्षण
जनगणना जनसंख्या के प्रत्येक सदस्य के बारे में जानकारी एकत्र करती है।	सर्वेक्षण जनसंख्या के नमूने से जानकारी एकत्र करता है
जनगणना अधिक विस्तृत और सटीक है	सर्वेक्षण जनगणना के रूप में सटीक या विश्वसनीय नहीं है
जनगणना को पूरा होने में लंबा समय लगता है	जनगणना की तुलना में सर्वेक्षण कम समय में किया जा सकता है।
जनगणना आम तौर पर सरकार द्वारा आयोजित की जाती है।	सर्वेक्षण किसी के द्वारा भी किया जा सकता है।
जनगणना अक्सर आयोजित नहीं की जाती है (हर 10 साल के बाद)	सर्वेक्षण अधिक बार आयोजित किए जा सकते हैं।

Data Processing

Data processing in research is the collection and translation of a data set into valuable, usable information. Through this process, a researcher, data engineer or data scientist takes raw data and converts it into a more readable format, such as a graph, report or chart, either manually or through an automated tool.

डाटा प्रासेसिंग

अनुसंधान में डेटा प्रोसेसिंग डेटा सेट का मूल्यवान, उपयोगी जानकारी में संग्रह और अनुवाद है। इस प्रक्रिया के माध्यम से, एक शोधकर्ता, डेटा इंजीनियर या डेटा वैज्ञानिक कच्चा डेटा लेता है और इसे मैनुअल रूप से या स्वचालित टूल के माध्यम से अधिक पठनीय प्रारूप, जैसे ग्राफ़, रिपोर्ट या चार्ट में परिवर्तित करता है।

Stages of Data Processing ✓

1. Data collection: Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality. ✓

डाटा प्रोसेसिंग के चरण

1. डेटा संग्रह: डेटा एकत्र करना डेटा प्रोसेसिंग में पहला कदम है। डेटा उपलब्ध स्रोतों से लिया जाता है, जिसमें डेटा लेक और डेटा वेयरहाउस शामिल हैं। यह महत्वपूर्ण है कि उपलब्ध डेटा स्रोत भरोसेमंद और अच्छी तरह से निर्मित हों ताकि एकत्र किया गया डेटा (और बाद में जानकारी के रूप में उपयोग किया जाए) उच्चतम संभव गुणवत्ता का हो।

1000 forms
970

2. Data preparation: Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as "pre-processing" is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data for the best business intelligence.

2. डेटा तैयारी: एक बार डेटा एकत्र हो जाने के बाद, यह डेटा तैयारी चरण में प्रवेश करता है। डेटा तैयारी, जिसे अक्सर "प्री-प्रोसेसिंग" के रूप में जाना जाता है, वह चरण है जिस पर कच्चे डेटा को साफ किया जाता है और डेटा प्रोसेसिंग के अगले चरण के लिए व्यवस्थित किया जाता है। तैयारी के दौरान, किसी भी त्रुटि के लिए कच्चे डेटा की सावधानीपूर्वक जाँच की जाती है। इस कदम का उद्देश्य खराब डेटा (अनावश्यक, अधूरा, या गलत डेटा) को खत्म करना और सर्वोत्तम व्यावसायिक बुद्धिमत्ता के लिए उच्च गुणवत्ता वाला डेटा बनाना शुरू करना है।

3. Data input: The clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.

3. डेटा इनपुट: साफ़ डेटा को उसके गंतव्य (शायद सेल्सफोर्स जैसा सीआरएम या रेडशिफ्ट जैसा डेटा वेयरहाउस) में दर्ज किया जाता है, और उस भाषा में अनुवाद किया जाता है जिसे वह समझ सकता है। डेटा इनपुट पहला चरण है जिसमें कच्चा डेटा उपयोगी जानकारी का रूप लेना शुरू कर देता है।

4. Processing: During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation.

Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed (data lakes, social networks, connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).

4. प्रसंस्करण: इस चरण के दौरान, पिछले चरण में कंप्यूटर में इनपुट किए गए डेटा को वास्तव में व्याख्या के लिए संसाधित किया जाता है। प्रसंस्करण मशीन लर्निंग एल्गोरिदम का उपयोग करके किया जाता है, हालांकि यह प्रक्रिया संसाधित किए जा रहे डेटा के स्रोत (डेटा झील, सोशल नेटवर्क, कनेक्टेड डिवाइस इत्यादि) और इसके इच्छित उपयोग (विज्ञापन पैटर्न की जांच, कनेक्टेड डिवाइस से चिकित्सा निदान) के आधार पर थोड़ी भिन्न हो सकती है। ग्राहकों की ज़रूरतों का निर्धारण, आदि)।

5. Data output/interpretation: The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

5. डेटा आउटपुट/व्याख्या: आउटपुट/व्याख्या चरण वह चरण है जिस पर डेटा अंततः गैर-डेटा वैज्ञानिकों के लिए उपयोग योग्य होता है। यह अनुवादित, पठनीय और अक्सर ग्राफ़, वीडियो, चित्र, सादे पाठ आदि के रूप में होता है। कंपनी या संस्थान के सदस्य अब अपने स्वयं के डेटा एनालिटिक्स प्रोजेक्ट के लिए डेटा की स्वयं-सेवा शुरू कर सकते हैं।

Input → processing → output

✓ 6. Data storage: The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. Plus, properly stored data is a necessity for compliance with data protection legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

General Data Protection Regulation

6. डेटा स्टोरेज: डेटा प्रोसेसिंग का अंतिम चरण स्टोरेज है। सभी डेटा संसाधित होने के बाद, इसे भविष्य में उपयोग के लिए संग्रहीत किया जाता है। हालाँकि कुछ जानकारी को तुरंत उपयोग में लाया जा सकता है, लेकिन इसमें से अधिकांश बाद में किसी उद्देश्य की पूर्ति करेगी। साथ ही, जीडीपीआर जैसे डेटा संरक्षण कानून के अनुपालन के लिए उचित रूप से संग्रहीत डेटा एक आवश्यकता है। जब डेटा ठीक से संग्रहीत किया जाता है, तो जरूरत पड़ने पर संगठन के सदस्यों द्वारा इसे जल्दी और आसानी से एक्सेस किया जा सकता है।

