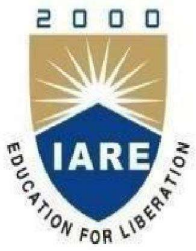
LECTURE NOTES

ON

DATA MANAGEMENET AND REPRESENTATION

Dr. M. V. Krisha Rao
Professor, CSE (DS)



COMPUTER SCIENCE AND ENGINEERING
INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
DUNDIGAL . 500 043, HYDERABAD

COURSE OBJECTIVES:

The students will try to learn:	
I	The data fundamentals, data collection, handling and preservation techniques.
II	The treatment of missed values in large data sets.
III	The data presentation and visual exploration techniques needed before the data analysis

COURSE OUTCOMES:

After successful completion of the course, Students will be able to:

CO No	Course Outcomes	Knowledge Level (Bloom's Taxonomy)
CO 1	Identify the data importing methods from the data files of various formats for data presentation and further exploration.	Remember
CO 2	Make use of imputation techniques for wrangling the data for subsequent data analysis.	Understand
CO 3	Identify the reasons of missing and bad data in various forms for applying cleaning techniques	Apply
CO 4	Examine different styles of tables and graphs for presenting and visualizing the data.	Analyze
CO 5	Determine the principles like clarity, precision and efficiency of data presentation and visualization	Evaluate
CO 6	Build different data visualizations using tabular data or dataframes.	Apply

MODULE-I

PRINCIPLES OF DATA MANAGEMENT

COURSE OUTCOMES:

CO No	Course Outcomes	Knowledge Level (Bloom's Taxonomy)
CO 1	Identify the data importing methods from the data files of various formats for data presentation and further exploration.	Understand

I. INTRODUCTION

Data is defined as facts or figures, or information that's stored in or used by a computer. An example of data is **information collected for a research paper**. An example of data is an email. ... Statistics or other information represented in a form suitable for processing by computer.

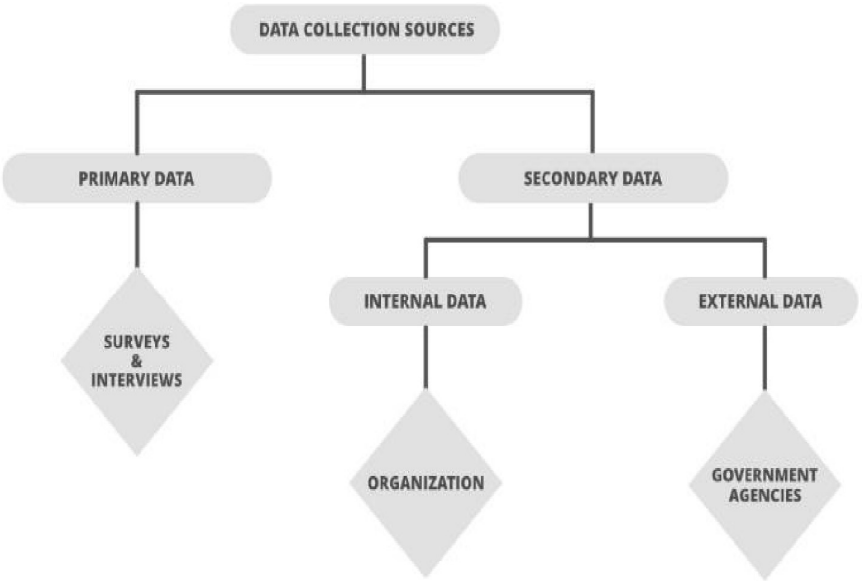
Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.

A Definition of Data Management

Data management is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users.

The actual data is then further divided mainly into two types known as:

1. **Primary data**
2. **Secondary data**



1. Primary data:

The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys. The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

1. Interview method:
The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:
The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

3. Observation method:
The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behavior towards the products. The data obtained will be sent for processing.

4. Experimental method:
The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.

- **CRD- Completely Randomized design** is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.
- **RBD- Randomized Block Design** is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.
- **LSD – Latin Square Design** is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.
- **FD- Factorial design** is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

2. Secondary data:

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

Internal source:
These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

External source:
The data which can't be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services, and other non-governmental publications.

Other sources:

- **Sensors data:** With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
- **Satellites data:** Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.
- **Web traffic:** Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.

3. CONCEPTS TO VARIABLES

What is a variable in data management?

A variable usually contains data from observations such as a score from a questionnaire, age or sex.

There are different types of variables and having their influence differently in a study viz. Independent & dependent variables, Active and attribute variables, Continuous, discrete and categorical variable, Extraneous variables and Demographic variables.

What are the 5 types of variables?

There are different types of variables and having their influence differently in a study viz. Independent & dependent variables, Active and attribute variables, Continuous, discrete and categorical variable, Extraneous variables and Demographic variables.

In scientific research, **concepts** are the abstract ideas or phenomena that are being studied (e.g., educational achievement). **Variables** are properties or characteristics of the concept (e.g., performance at school), while **indicators** are ways of measuring or quantifying variables (e.g., yearly grade reports).

Can concepts be converted into variables?
Concepts should be converted into variables so that they can be measured, although on different scales same variable will have different precision. If the researcher is using some concepts in his research he needs to find out some indicators that are reflective of these concepts.

4. WHAT ARE THE DIFFERENT FORMS OF DATA?

What is Data?

- Text (e.g. doc, txt, pdf)
- Numerical (e.g. SPSS, STATA, . xls, Access, MySQL)
- Multimedia (e.g. jpeg, tiff, wav, mpeg, quicktime)
- Models (e.g. 3D, statistical)
- Software (e.g. Java, C)
- Domain-specific (e.g. FITS in astronomy, CIF in chemistry)
- Instrument-specific (e.g. Olympus Confocal Microscope Data Format)

4 Types of Data: **Nominal, Ordinal, Discrete, Continuous**. What are the 7 types of data?

- Useless.
- Nominal.
- Binary.
- Ordinal.
- Count.
- Time.
- Interval.

5. WHAT IS A CODEBOOK?

A codebook describes the contents, structure, and layout of a data collection. A well-documented codebook "contains information intended to be complete and self-explanatory for each variable in a data file[1]."

Codebooks begin with basic front matter, including the study title, name of the principal investigator(s), table of contents, and an introduction describing the purpose and format of the codebook. Some codebooks also include methodological details, such as how weights were computed, and data collection instruments, while others, especially with larger or more complex data collections, leave those details for a separate user guide and/or data collection instrument.

The main body of a codebook contains unambiguous variable level details.

```
H00034.00     [H40-SF12-2]                                    Survey Year: 2002

              SF12 - ASSESSMENT OF R'S GENERAL HEALTH

In general, would you say your health is ....

NOTE: SF-12(r) Health Survey (Medical Outcomes Trust)
(c) Medical Outcomes Trust and John E. Ware, Jr., All Rights Reserved
SF-12(tm) (QualityMetric, Inc.)

     1232        1 Excellent
     2111        2 Very Good
     1531        3 Good
      563        4 Fair
      145        5 Poor
   -------
     5582

Refusal(-1)              6
Don't Know(-2)           0
TOTAL ---------->     5588    VALID SKIP(-4)     7098      NON-INTERVIEW(-5)        0

Lead In: H00033.00[Default]
Default Next Question: H00035.00
```

- **Variable name:** The name or number assigned to each variable in the data collection. Some researchers prefer to use mnemonic abbreviations (e.g., EMPLOY1), while others use alphanumeric patterns (e.g., VAR001). For survey data, try to name variables after the question numbers - e.g., Q1, Q2b, etc.
- **Variable label:** A brief description to identify the variable for the user.
- **Question text:** Where applicable, the exact wording from survey questions. ["In general, would you say your health is . . ."]
- **Values:** The actual coded values in the data for this variable.
- **Value labels:** The textual descriptions of the codes
- **Summary statistics:** Where appropriate and depending on the type of variable, provide unweighted summary statistics for quick reference. For categorical variables, for instance, frequency counts showing the number of times a value occur and the percentage of cases that value represents for the variable are appropriate. For continuous variables, minimum, maximum, and median values are relevant.
- **Missing data:** Where applicable, the values and labels of missing data. Missing data can bias an analysis and is important to convey in study documentation. Remember to describe all missing codes, including "system missing" and blank.
- **Universe skip patterns:** Where applicable, information about the population to which the variable refers, as well as the preceding and following variables.
- **Notes:** Additional notes, remarks, or comments that contextualize the information conveyed in the variable or relay special instructions. For measures or questions from copyrighted instruments, the notes field is the appropriate location to cite the source.

6. WHAT IS DATA DOCUMENTATION?

Documenting your data is simply **providing sufficient descriptive information about your data** so that it can be used properly by you, your colleagues, and other researchers in the future. Well documented data is identifiable, understandable, and usable in the future.

What is documentation explain?
Documentation is **any communicable material that is used to describe, explain or instruct regarding some attributes of an object**, system or procedure, such as its parts, assembly, installation, maintenance and use. ... Documentation is often distributed via websites, software products, and other online applications.

What is the importance of documentation in data management?
Data documentation will **ensure that your data will be understood and interpreted by any user**. It will explain how your data was created, what the context is for the data, structure of the data and its contents, and any manipulations that have been done to the data.

7. WHAT IS DATA MANAGEMENT CODING?

Coding of data refers to **the process of transforming collected information or observations to a set of meaningful, cohesive categories**. It is a process of summarizing and re-presenting data in order to provide a systematic account of the recorded or observed phenomenon.

What are the types of coding?
There are four types of coding:

- Data compression (or source coding)
- Error control (or channel coding)
- Cryptographic coding.
- Line coding.

Who uses coding?

9 Computer coding and programming jobs to consider

- Software application developer.
- Web developer.
- Computer systems engineer.
- Database administrator.
- Computer systems analyst.
- Software quality assurance (QA) engineer.
- Business intelligence analyst.
- Computer programmer. **8. WHAT IS DATA**

CLEANING AND SCREENING?

- Data cleaning and screening is **the step that directly follows data entry and you must not start your analysis unless doing it**. 4 Data screening importance: -1 It is very easy to make mistakes when entering data. -1 Some errors can miss up your analysis.

What is meant by data cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

What is data cleaning in data analysis?

Data Cleaning means the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according to the necessity. Data cleaning is considered a foundational element of the basic data science.

What is data screening?

Data screening (sometimes referred to as "data screaming") is the process of ensuring your data is clean and ready to go before you conduct further statistical analyses. Data must be screened in order to ensure the data is useable, reliable, and valid for testing causal theory.

What are the steps of data screening?

Data Screening Short Course

1. Entering and checking raw data.
2. Assessing univariate problems (distribution shape, outliers)
3. Assessing bivariate problems (linearity, regression diagnostics)
4. Assessing multivariate problems (multivariate normality, detecting multivariate outliers)
5. Dealing with missing data.

Why is data screening necessary?

Data screening should be conducted prior to data recoding and data analysis, to **help ensure the integrity of the data**. It is only necessary to screen the data for the variables and cases used for the analyses presented in the lab report. Data screening means checking data for errors and fixing or removing these errors.

9. WHAT ARE THE FIVE FUNCTIONS OF FILE MANAGEMENT?

What are the basic functions of file management in OS?

- Creating. It helps in creating a new file at the specified location in a computer system.
- Saving. It helps in saving the content written in a file at some specified location.
- Opening.
- Modifying.
- Closing.
- Renaming.
- Deleting.

What are the 3 basic types of file management?
There are three basic types of special files: **FIFO (first-in, first-out), block, and character**. FIFO files are also called pipes. Pipes are created by one process to temporarily allow communication with another process. These files cease to exist when the first process finishes.

What are file management strategies?
These file management tips will help you keep your files accessible:

- Organize by file types.
- One place for all.
- Create folders in My Documents.
- Nest folders within folders.
- Follow the file naming conventions.
- Be specific.
- File as you go.
- Order your files for your convenience.

What are the key principles of data management?

6 key data management principles

- Create a data management strategy. One of the most important data management principles is developing a data management plan. ...
- Define roles in the data management system. ...
- Control data throughout its life cycle. ...
- Ensure data quality. ...
- Collect and analyze metadata. ...
- Maximize the use of data.

MODULE II

SECONDARY, PRIMARY AND ADMINISTRATIVE DATA

COURSE OUTCOMES:

CO No	Course Outcomes	Knowledge Level (Bloom's Taxonomy)
CO 1	Identify the data importing methods from the data files of various formats for data presentation and further exploration.	Remember

1. SECONDARY DATA

Secondary data refers to **data that is collected by someone other than the primary user**. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data that was originally collected for other research purposes.

2. WHAT ARE THE TYPES OF SECONDARY DATA?

There are two common types of secondary data: **Internal data and External data**. Internal data is the information that has been stored or organized by the organization itself. External data is the data organized or collected by someone else.

What Are The Four Types Of Secondary Data?

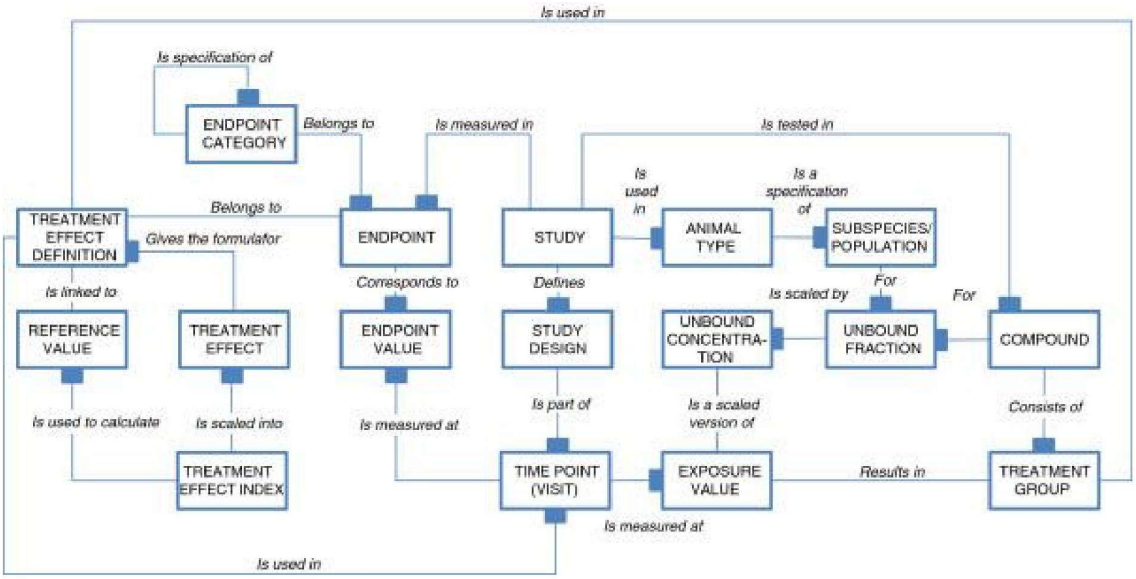
Sources of secondary data include **books, personal sources, journals, newspapers, websitess, government records etc**. Secondary data are known to be readily available compared to that of primary data. It requires very little research and needs for manpower to use these sources.

3. WHAT ARE THE USES OF SECONDARY DATA? Uses of secondary data

- Identify the research problem.
- Develop a strategy to arrive at solutions to the problem.
- Develop a strategy to arrive at solutions to the problem.
- Formulate an appropriate research design.
- Find the answers to certain research questions or test some hypotheses.
- Interpret primary data.

4. WHAT IS A CONCEPTUAL DATA MODEL IN INFORMATION MANAGEMENT?

A conceptual data model is **developed based on the data requirements for the application that is being developed**, perhaps in the context of an activity model. The data model will normally consist of entity types, attributes, relationships, integrity rules, and the definitions of those objects.


Is used in
Is specification of
ENDPOINT CATEGORY
Belongs to
Is measured in
Is tested in
Is used in
Is a specification of
TREATMENT EFFECT DEFINITION
Belongs to
Gives the formulator
ENDPOINT
STUDY
ANIMAL TYPE
SUBSPECIES/ POPULATION
Is linked to
Corresponds to
Defines
Is scaled by
For
For
REFERENCE VALUE
TREATMENT EFFECT
ENDPOINT VALUE
STUDY DESIGN
UNBOUND CONCENTRA-TION
UNBOUND FRACTION
COMPOUND
Is used to calculate
Is scaled into
Is measured at
Is part of
Is a scaled version of
Consists of
TREATMENT EFFECT INDEX
TIME POINT (VISIT)
EXPOSURE VALUE
Results in
TREATMENT GROUP
Is measured at
Is used in

6. WHAT IS ADMINISTRATIVE DATA?

Administrative data is the data that organizations collect about their operations. It includes data for routine operations, and is frequently used to assess how well an organization is achieving its intended goals.

What are administrative data elements?

Border records, pensions, taxation, and vital records like births and deaths are examples of administrative data. ... These types of data are used to produce management information, like registration data in a cost-effective way.

What is administrative data in a health record?

Administrative data include enrollment or eligibility information, claims information, and managed care encounters. ... The claims and encounters may be for hospital and other facility services, professional services, prescription drug services, laboratory services, and so on.

7. DATA LINKING

Data linking is used to bring together information from different sources in order to create a new, richer dataset. This involves identifying and combining information from corresponding records on each of the different source datasets.

What is linking data sets?

Data linking is used to bring together information from different sources in order to create a new, richer dataset. This involves identifying and combining information from corresponding records on each of the different source datasets.

MODULE III

SECONDARY, PRIMARY AND ADMINISTRATIVE DATA COURSE OUTCOMES

CO No	Course Outcomes	Knowledge Level (Bloom's Taxonomy)
CO 2	Make use of imputation techniques for wrangling the data for subsequent data analysis.	Understand
CO 3	Identify the reasons of missing and bad data in various forms for applying cleaning techniques	Apply

1. WHY ARE MISSING DATA A PROBLEM?

Missing data **can cause serious problems**. ... This means that in the end, you may not have enough data to perform the analysis. For example, you could not run a factor analysis on just a few cases. Second, the analysis might run but the results may not be statistically significant because of the small amount of input data.

2. REASONS FOR MISSING DATA

Many reasons for missing data ...

- People do not respond to survey (or specific questions in a survey).
- Species are rare and cannot be found or sampled.
- The individual dies or drops out before sampling.
- Some things are easier to measure than others.
- Data entry errors.
- Many others!

Why does missing data happen?
In statistics, missing data, or missing values, occur **when no data value is stored for the variable in an observation**. ... Sometimes missing values are caused by the researcher for example, when data collection is done improperly or mistakes are made in data entry.

3. WHAT ARE THE THREE TYPES OF MISSING DATA?

Missing data are typically grouped into three categories:

~ Missing completely at random (MCAR). When data are MCAR, the fact that the data are missing is independent of the observed and unobserved data. ...

- Missing at random (MAR). ...
- Missing not at random (MNAR). **Forms**

or Types of Missing Data

When considering the potential impact of the missing data on the registry findings, it is important to consider the underlying reasons for why the data are missing.[14] Missing data are typically grouped into three categories:

~ Missing completely at random (MCAR). When data are MCAR, the fact that the data are missing is independent of the observed and unobserved data.[15] In other words, no systematic differences exist between participants with missing data and those with complete data. For example, some participants may have missing laboratory values because a batch of lab samples was processed improperly. In these instances, the missing data reduce the analyzable population of the study

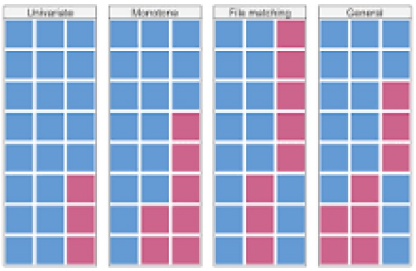
and consequently, the statistical power, but do not introduce bias: when data are MCAR, the data which remain can be considered a simple random sample of the full data set of interest. MCAR is generally regarded as a strong and often unrealistic assumption.

- Missing at random (MAR). When data are MAR, the fact that the data are missing is systematically related to the observed but not the unobserved data.[15] For example, a registry examining depression may encounter data that are MAR if male participants are less likely to complete a survey about depression severity than female participants. That is, if probability of completion of the survey is related to their sex (which is fully observed) but not the severity of their depression, then the data may be regarded as MAR. Complete case analyses, which are based on only observations for which all relevant data are present and no fields are missing, of a data set containing MAR data may or may not result in bias. If the complete case analysis is biased, however, proper accounting for the known factors (in the above example, sex) can produce unbiased results in analysis.

- Missing not at random (MNAR). When data are MNAR, the fact that the data are missing is systematically related to the unobserved data, that is, the missingness is related to events or factors which are not measured by the researcher. To extend the previous example, the depression registry may encounter data that are MNAR if participants with severe depression are more likely to refuse to complete the survey about depression severity. As with MAR data, complete case analysis of a data set containing MNAR data may or may not result in bias; if the complete case analysis is biased, however, the fact that the sources of missing data are themselves unmeasured means that (in general) this issue cannot be addressed in analysis and the estimate of effect will likely be biased.

4. PATTERNS OF MISSING DATA

A missing data pattern is said to be **univariate if there is only one variable with missing data**. ... A missing data pattern is said to be monotone if the variables Yj can be ordered such that if Yj is missing then all variables Yk with k>j are also missing. This occurs, for example, in longitudinal studies with drop-out.



Missing data (MD) is a prevalent problem and can negatively affect the trustworthiness of data analysis. In industrial use cases, faulty sensors or errors during data integration are common causes for systematically missing values.
The majority of MD research deals with imputation, i.e., the replacement of missing values with “best guesses”. Most imputation methods require missing values to occur independently, which is rarely the case in industry. Thus, it is necessary to identify missing data patterns (i.e., systematically missing values) prior to imputation (1) to understand the cause of the missingness, (2) to gain deeper insight into the data, and (3) to choose the proper imputation technique. However, in literature, there is a wide varity of MD patterns without a common formalization.

5. ADDRESSING MISSING DATA IN THE ANALYSIS STAGE

How do you handle missing data in data analysis?

Best techniques to handle missing data

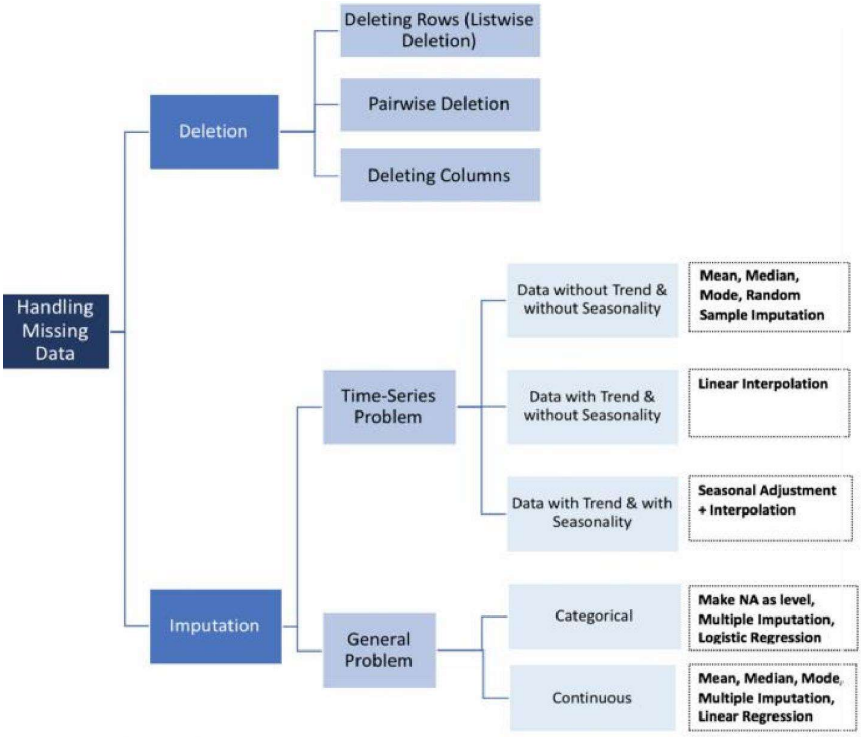
1. Use deletion methods to eliminate missing data. The deletion methods only work for certain datasets where participants have missing fields. ...
2. Use regression analysis to systematically eliminate data. ...
3. Data scientists can use data imputation techniques.

How do you address missing data in research?

By far the most common approach to the missing data is to simply omit those cases with the missing data and analyze the remaining data. This approach is known as the **complete case (or available case) analysis or listwise deletion**.

How do you handle missing or corrupted data in a dataset?

1. Method 1 is deleting rows or columns. We usually use this method when it comes to empty cells. ..
2. Method 2 is replacing the missing data with aggregated values. ...
3. Method 3 is creating an unknown category. ...
4. Method 4 is predicting missing values.



MODULE IV

DATA PRESENTATION

COURSE OUTCOMES:

CO No	Course Outcomes	Knowledge Level (Bloom's Taxonomy)
CO 4	Examine different styles of tables and graphs for presenting and visualizing the data.	Analyze
CO 5	Determine the principles like clarity, precision and efficiency of data presentation and visualization	Evaluate

1. PRESENTING DATA

What should be considered when presenting data?

- 1) Make sure your data can be seen. ...
- 2) Focus most on the points your data illustrates. ...
- 3) Share one — and only one — major point from each chart. ...
- 4) Label chart components clearly. ...
- 5) Visually highlight "Aha!" zones. ...
- 6) Write a slide title that reinforces the data's point. ...
- 7) Present to your audience, not to your data.

How to present data visually (data visualization best practices)

1. Avoid distorting the data. ...
2. Avoid cluttering up your design with "chartjunk" ...
3. Tell a story with your data. ...
4. Combine different types of data visualizations. ...
5. Use icons to emphasize important points. ...
6. Use bold fonts to make text information engaging.

What are the types of presenting data?

The three main forms of presentation of data are:

- Textual presentation.
- Data tables.
- Diagrammatic presentation.

How many methods there are in presenting data?

Data is initially collected from a given source, whether they are experiments, surveys, or observation, and is presented in one of **four methods**: Textual Method. The reader acquires information through reading the gathered data. Tabular Method.

What is the most effective method of presenting data?

While **graphs** are effective for presenting large amounts of data, they can be used in place of tables to present small sets of data. A graph format that best presents information must be chosen so that readers and reviewers can easily understand the information.

What are data presentation tools?
Data tools include **standard charts and graphs**, such as a bar chart, block histogram, bubble chart, scatterplot, pie chart, line graph, and so on. Users can also choose to display data as networks of related words and ideas, such as a word tree, tag cloud, or word cloud.

2. VISUAL IMAGES

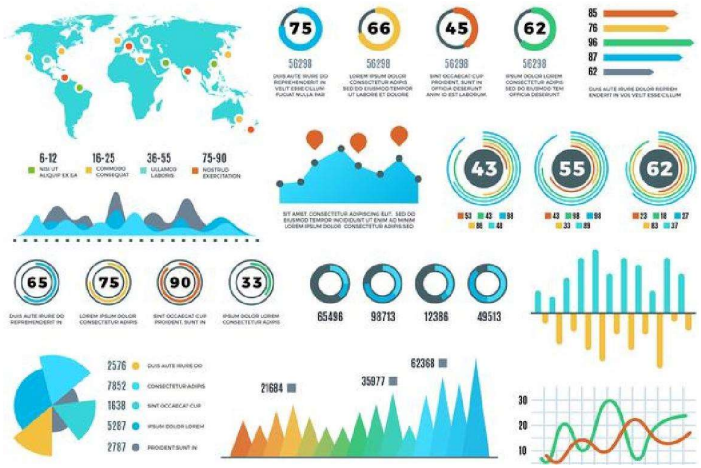
What is a visual way to show data?
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Why are visual representations important in presenting data?
Data visualization gives us a clear idea of what the information means by **giving it visual context through maps or graphs**. This makes the data more natural for the human mind to comprehend and therefore makes it easier to identify trends, patterns, and outliers within large data sets.

What are five formats for the visual display of data?

1. Column Chart. This is one of the most common types of data visualization tools. ...
2. Bar Graph. ...
3. Stacked Bar Graph. ...
4. Line Graph. ...
5. Dual-Axis Chart. ...
6. Mekko Chart. ...
7. Pie Chart. ...
8. Scatter Plot.

What is data visualization with examples?
"Data visualization"refers to **transforming figures and raw data into visual objects**: points, bars," line plots, maps, etc. By combining user-friendly and aesthetically pleasing features, these visualizations make research and data analysis much quicker and are also a powerful communication tool.



What are the principles of data presentation?
Again, the principle that should guide decisions is that it is the data which must be presented.
...
Friendly scatter graph

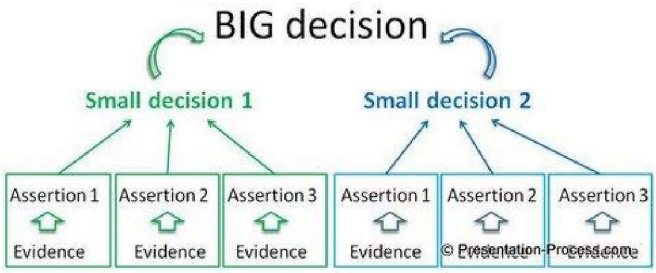
- Above all else show the data.
- Maximize the data-ink ratio.
- Erase non-data ink (within reason)
- Erase redundant data ink.
- Revise and re-edit.

3. WHAT IS THE CLARITY OF PRESENTATION?

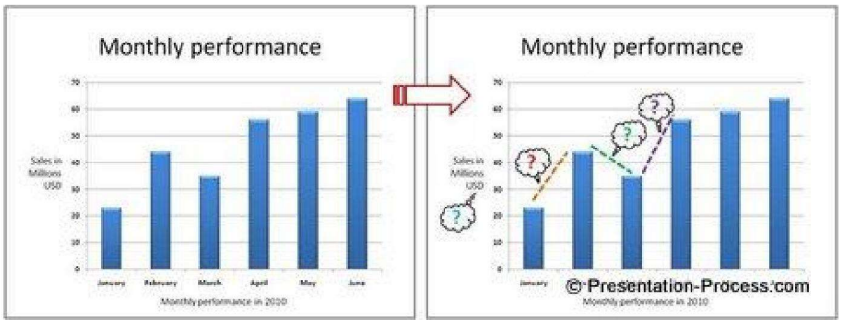
Clarity is defined as "clearness or lucidity as to perception or understanding; freedom from indistinctness or ambiguity." It's the latter part that really sticks out to us as a virtue of presentation design. "Freedom from indistinctness or ambiguity," is precisely what we need to achieve in our presentations.

Setting the context for clarity in presentation of data:

Every element in a presentation should lead the audience to make the BIG decision. Here is a visual representation of the way the various elements in a presentation lead to the BIG decision:



Your presentation will be successful, when you make the path towards the BIG decision as smooth and clear as possible for your audience. For this, you need to make the inference from your slides obvious.



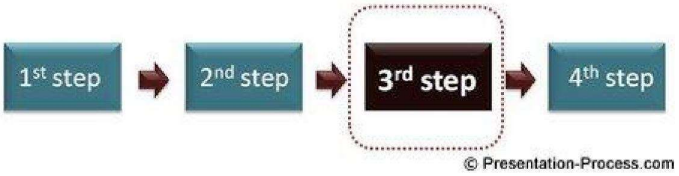
Though the chart gives the right information, the inference from the slide is not obvious. When there is no clarity in presentation of data, different people in the audience start looking at different parts of the chart and raise questions. The discussion goes off on a tangent and the presenter loses control of the situation.

Consider this alternative representation of data in the graph:

This slide title clearly spells out the message of the slide. The specific portion of the chart which proves the assertion is clearly highlighted using a green dotted arrow. The audience 'gets' the message in seconds and moves towards the BIG decision.

How do you bring in clarity in presentation of data?

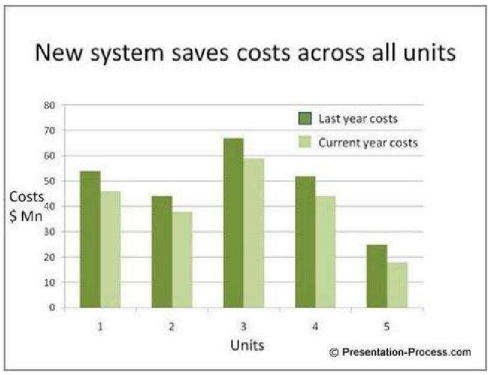
The first way is to use **contrast** to highlight the important information.
You can choose to use a different color to bring attention to the important point of your slide, or highlight the main point with a red line or use any method that makes the audience notice the information you want them to register in their mind.



Let us see how we can enhance the clarity of a data presentation slide:

In the following slide, the presenter wants to showcase that their new system has saved costs across all units.
So, the right data to show is the cost comparison between previous year (when the system was not available) and the current year (with the new system).
The information is captured in the following slide:

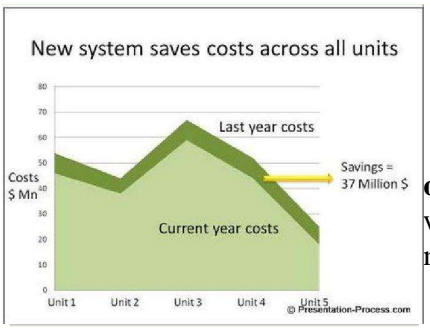
Though the two charts used in the slide provide the required data, the inference is not obvious. Here is a better alternative:



In this, both the charts of the previous slide are combined to help the audience read the information easily. This definitely improves the clarity of the information provided. But, the inference is not made obvious enough.

Here is a much better alternative of presenting the graph with clarity

The information is provided in a clearly understandable form. The exact amount saved is mentioned on the slide in a very obvious way.



4. PRECISION

What is precision example?
Precision refers **to the closeness of two or more measurements to each other**. Using the example above, if you weigh a given substance five times, and get 3.2 kg each time, then your measurement is very precise. Precision is independent of accuracy.

What does precision mean in data?
Precision is **the degree to which future measurements or calculations yield the same or similar results** it is a measure of the spread of repeated measurement results and depends only on the distribution of random errors it gives no indication of how close those results are to the true value.

How do you find the precision of data?

In an imbalanced classification problem with two classes, precision is calculated as **the number of true positives divided by the total number of true positives and false positives**. The result is a value between 0.0 for no precision and 1.0 for full or perfect precision.

Important Techniques for Efficient Data Presentation and Insights

1 - Create Group, Hierarchy, Sets, use filter efficiently and keep it simple
2 - Develop a better sense of "graphicacy."

5. TYPES OF TABLES

What is a table in data presentation?

A Table refers **to any data which is presented in orderly rows across and/or down the page, often enclosed within borders**. A Figure refers to any other form of presentation such as a bar or pie chart, a graph, a diagram, a map, a photograph, a line drawing or a sample of material.

What are the different types of data tables?
There are three types of tables: **base, view, and merged**. Every table is a document with its own title, viewers, saved visualizations, and set of data.

What are the types of data presentation?
The three main forms of presentation of data are:

- Textual presentation.
- Data tables.
- Diagrammatic presentation.

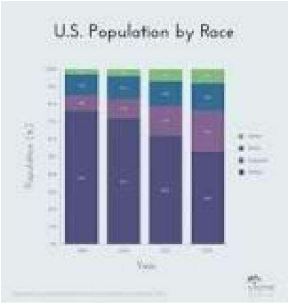
6. WHAT IS GRAPHICAL PRESENTATION OF DATA?

Data is ingested into graphical representation of data software and then represented by a variety of symbols, such as lines on a line chart, bars on a bar chart, or slices on a pie chart, from which users can gain greater insight than by numerical analysis alone.

Some of the various types of graphical representation include:

- Line Graphs.
- Bar Graphs.
- Histograms.
- Line Plots.
- Frequency Table.
- Circle Graph, etc.

What are the types of data graphs?
Popular graph types include line graphs, bar graphs, pie charts, scatter plots and histograms. Graphs are a great way to visualize data and display statistics. For example, a bar graph or chart is used to display numerical data that is independent of one another.



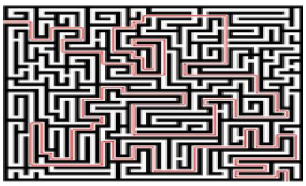
7. PRINCIPLES OF DATA PRESENTATION.

Again, the principle that should guide decisions is that it is the data which must be presented. ...
Friendly scatter graph

- Above all else show the data.
- Maximize the data-ink ratio.
- Erase non-data ink (within reason)
- Erase redundant data ink.
- Revise and re-edit.

What are the steps in data presentation?

1. Steps in Preparing a Presentation.
2. Planning Your Presentation.
3. Step 1: Analyze your audience.
4. Step 2: Select a topic.
5. Step 3: Define the objective of the presentation.
6. Preparing the Content of Your Presentation.
7. Step 4: Prepare the body of the presentation.
8. Step 5: Prepare the introduction and conclusion.



What are the four methods of data presentation?
In this article, the techniques of data and information presentation in textual, tabular, and graphical forms are introduced. Text is the principal method for explaining findings, outlining trends, and providing contextual information.

MODULE V

DESIGNING TABLES AND GRAPHICS FOR DATA PRESENTATIONS

COURSE OUTCOMES

CO No	Course Outcomes	Knowledge Level (Bloom's Taxonomy)
CO 6	Build different data visualizations using tabular data or dataframes.	Apply

1. TABLES

Is table a graphic?

What are tables and graphs? Tables and graphs are **visual representations**. They are used to organize information to show patterns and relationships. A graph shows this information by representing it as a shape.

What is a table in data presentation?

A Table refers to any data which is presented in orderly rows across and/or down the page, often enclosed within borders. A Figure refers to any other form of presentation such as a bar or pie chart, a graph, a diagram, a map, a photograph, a line drawing or a sample of material.

What are tables used for in a presentation?

Tables are another tool you can use **to display information in PowerPoint**. A table is a grid of cells arranged in rows and columns. Tables are useful for various tasks, including presenting text information and numerical data. You can even customize tables to fit your presentation.

A graphical table is **a summarizing visualization designed to provide a lot of information at one glance**. It can be set up to show columns with dynamic items such as sparklines, calculated values, conditional icons, or bullet graphs. One value is shown for each row as specified on the Rows axis

How do you present data from a table? Presenting data in tables

1. Preparation of tables.
2. Title. Every table must have a brief descriptive title. ...
3. Structure. ...
4. Headings and sub-headings. ...
5. Numerical data. ...
6. Other notations. ...
7. Statistics. ...
8. Text.

What types of data are represented in tables?

Tables show quantitative data effectively. They may be used to communicate precise magnitudes.

What is Tabular Presentation of Data?

It is a table that helps to represent even a large amount of data in an engaging, easy to read, and coordinated manner. The data is arranged in rows and columns. This is one of the most popularly used forms of presentation of data as data tables are simple to prepare and read.

The most significant benefit of tabulation is that it coordinates data for additional statistical treatment and decision making. The analysis used in tabulation is of four types. They are:

1. Qualitative
2. Quantitative
3. Temporal
4. Spatial

1. Qualitative classification: When the classification is done according to traits such as physical status, nationality, social status, etc., it is known as qualitative classification.

2. Quantitative classification: In this, the data is classified on the basis of features that are quantitative in nature. In other words, these features can be estimated quantitatively.

3. Temporal classification: In this classification, time becomes the categorizing variable and data are classified according to time. Time, maybe in years, months, weeks, days, hours, etc.,

4. Spatial classification: When the categorization is done on the basis of location, it is known as spatial classification. The place may be a country, state, district, block, village/town, etc.

Basics of Tabular Presentation

- Tabulation, i.e., tabular presentation of data is a method of presentation of data.
- It is a systematic and logical arrangement of data in the form of rows and columns with respect to the characteristics of data.
- It is an orderly arrangement which is compact and self-explanatory.
- Its objective is to: Present the data in a simple form, economies (save) space, facilitate comparison, facilitate statistical analysis, reduce the chances of errors

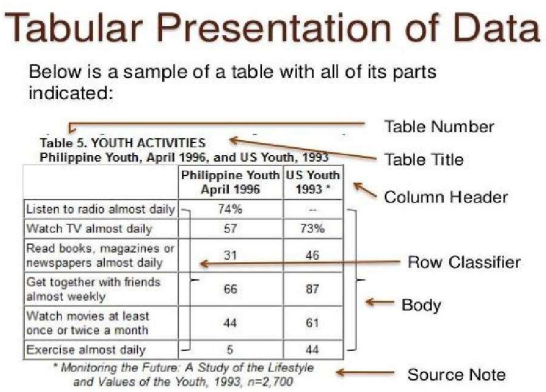
Main Parts of a Table:

Following are the main parts of a table:

(1) Table number	• Table number is the very first item mentioned on the top of each table for easy identification and further reference.
(2) Title	• Title of the table is the second item that is shown just above the table. • It narrates the contents of the table, hence it has to be very clear, brief, and carefully worded.
(3) Head note	• It is the third item just above the table and shown after the title. • It gives information about units of data like, 'amount in rupees or $', "quantity in tonnes', etc. • It is generally given in brackets.
(4) Captions or Column headings	• At the top of each column in a table, a column designation/head is given to explain the figures of the column. • This column heading is known as 'caption'.

(5) Stubs or Row headings	• The title of the horizontal rows is known as 'stubs'.
(6) Body of the table	• It contains the numeric information and reveals the whole story of investigated facts. Columns are read vertically from top to bottom and rows are read horizontally from left to right.
(7) Source note	• It is a brief statement or phrase indicating the source of data presented in the table.
(8) Footnote	• It explains the specific feature of the table which is not self-explanatory and has not been explained earlier. For example, points of exception if any.

2. EXAMPLES OF TABLES



State/UT	Population			Number of literates			Literacy Rate		
	Total (millions)	Males (millions)	Females (millions)	Total literates (millions)	Males (millions)	Females (millions)	Total (%)	M (%)	F (%)
J & K	10.1	5.4	4.8	4.8	3.1	1.7	55.5	66.6	43.0
HP	6.0	3.1	3.0	4.0	2.3	1.8	76.5	85.3	67.4
Punjab	24.3	13.0	11.4	14.8	8.4	6.3	69.7	75.2	63.4
Chandigarh	0.9	0.5	4.0	0.6	3.8	0.3	81.9	86.1	76.5
Uttaranchal	8.5	4.3	4.2	5.1	3.0	2.1	71.6	83.3	59.6
Haryana	21.1	11.4	9.8	12.1	7.5	4.6	67.9	78.5	55.7
Delhi	13.9	7.6	6.2	9.7	5.7	3.0	81.7	87.3	74.7
Rajasthan	56.5	29.4	27.1	27.7	18.0	9.7	60.4	75.7	43.9

3. GRAPHICS

What is graphic presentation of data?

Graphical representation refers to the **use of charts and graphs** to visually display, analyze, clarify, and interpret numerical data, functions, and other qualitative structures.

What are the different methods of graphical presentation of data?

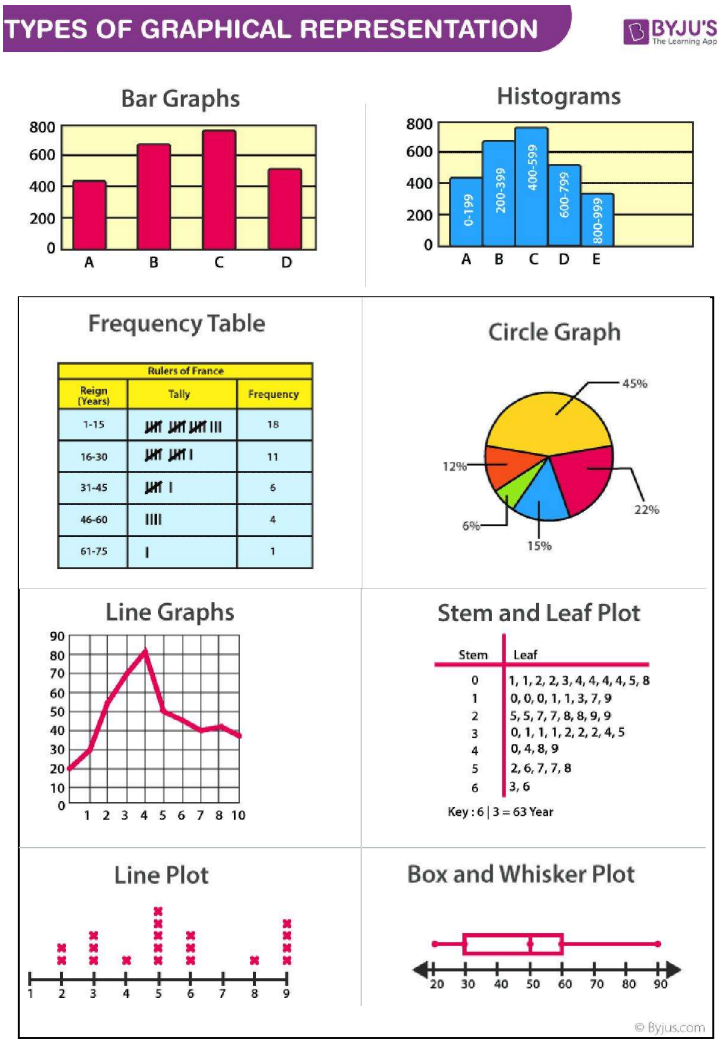
Generally, four methods are used to represent a frequency distribution graphically. These are Histogram, Smoothed frequency graph and Ogive or Cumulative frequency graph and pie diagram.

Some of the various types of graphical representation include:

- Line Graphs.
- Bar Graphs.

- Histograms.
- Line Plots.
- Frequency Table.
- Circle Graph, etc.

4. EXAMPLES OF GRAPHICS



5. EXAMPLES OF DESIGNING TABLES AND GRAPHS FOR DATA PRESENTATION

Variables - are constituted by data. For instance, an individual may be male or female. In this case, there are 10 observations for each sex, but "sex" is the variable that is referred to as a whole. Another example of variable is "age" in complete years, in which observations are the values 1 year, 2 years, 3 years, and so forth. In other words, variables are characteristics or attributes that can be measured, assuming different values, such as sex, skin type, eye color, age of the individuals under study, laboratory results, or the presence of a given lesion/disease. Variables are specifically divided into two large groups: **(a)** the group of categorical or qualitative variables, which is subdivided into dichotomous, nominal and ordinal variables; and **(b)** the group of numerical or quantitative variables, which is subdivided into continuous and discrete variables.

Categorical variables

a. Dichotomous variables, also known as binary variables: are those that have only two categories, i.e., only two response options. Typical examples of this type of variable are sex (male and female) and presence of skin cancer (yes or no).

b. Ordinal variables: are those that have three or more categories with an obvious ordering of the categories (whether in an ascending or descending order). For example, Fitzpatrick skin classification into types I, II, III, IV and V.[1]

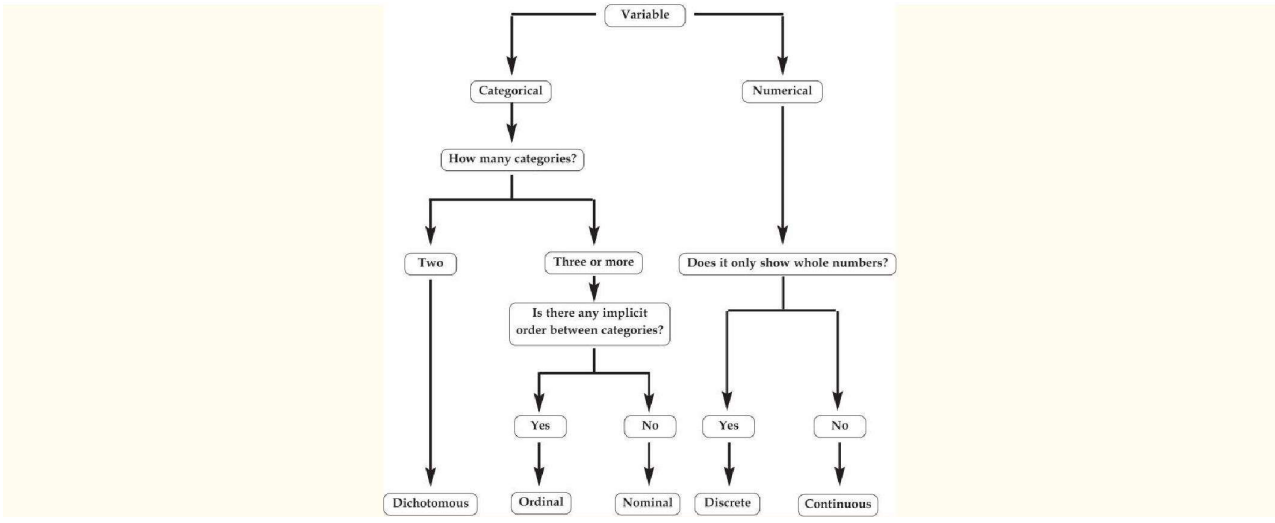
c. Nominal variables: are those that have three or more categories with no apparent ordering of the categories. Example: blood types A, B, AB, and O, or brown, blue or green eye colors.

Numerical variables

a. Discrete variables: are observations that can only take certain numerical values. An example of this type of variable is subjects' age, when assessed in complete years of life (1 year, 2 years, 3 years, 4 years, etc.) and the number of times a set of patients visited the dermatologist in a year.

b. Continuous variables: are those measured on a continuous scale, i.e., which have as many decimal places as the measuring instrument can record. For instance: blood pressure, birth weight, height, or even age, when measured on a continuous scale.

It is important to point out that, depending on the objectives of the study, data may be collected as discrete or continuous variables and be subsequently transformed into categorical variables to suit the purpose of the research and/or make interpretation easier. However, it is important to emphasize that variables measured on a numerical scale (whether discrete or continuous) are richer in information and should be preferred for statistical analyses. Figure 1 shows a diagram that makes it easier to understand, identify and classify the abovementioned variables.



Firstly, it is worth emphasizing that every table or graph should be self-explanatory, i.e., should be understandable without the need to read the text that refers to it refers.

a)Presentation of categorical variables

In order to analyze the distribution of a variable, data should be organized according to the occurrence of different results in each category. As for categorical variables, frequency distributions may be presented in a table or a graph, including bar charts and pie or sector charts. The term *frequency distribution* has a specific meaning, referring to the way observations of a given variable behave in terms of its absolute, relative or cumulative frequencies.

In order to synthesize information contained in a categorical variable using a table, it is important to count the number of observations in each category of the variable, thus obtaining its absolute frequencies. However, in addition to absolute frequencies, it is worth presenting its percentage values, also known as relative frequencies. For example, table 1 expresses, in absolute and relative terms, the frequency of acne scars in 18-year-old youngsters from a population-based study conducted in the city of Pelotas, Southern Brazil, in 2010.[3]

TABLE 1

Absolute and relative frequencies of acne scar in 18- year-old adolescents (n = 2.414). Pelotas, Brazil, 2010

Prevalence	Absolute frequency (n)	Relative frequency (%)
No	1.855	76.84
Yes	559	23.16
Total	2.414	100.00

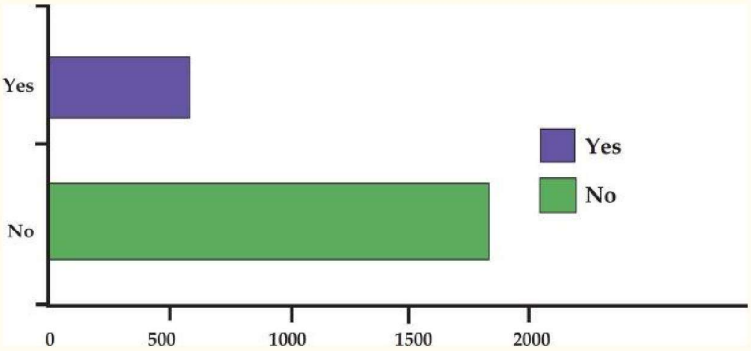


FIGURE 2

Absolute frequencies of acne scar in 18-year-old adolescents (n = 2.414). Pelotas, Brazil, 2010

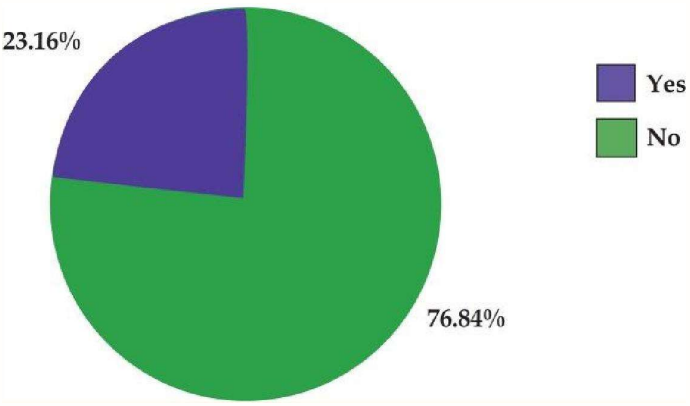


FIGURE 3

Relative frequencies of acne scar in 18-year-old adolescents (n = 2.414). Pelotas, Brazil, 2010 ***b)Presentation of numerical variables***

Frequency distributions of numerical variables can be displayed in a table, a histogram chart, or a frequency polygon chart. With regard to discrete variables, it is possible to present the number of observations according to the different values found in the study, as illustrated in table 2. This type of table may provide a wide range of information on the collected data.

TABLE 2

Educational level of 18-year-old adolescents (n = 2,199). Pelotas, Brazil, 2010

Educational level (in years of education)	Absolute frequency (n)	Relative frequency (%)	Cumulative relative frequency (%)
Total	**2.199**	**100.00**	**-**
0	1	0.05	0.05
1	2	0.09	0.14
2	2	0.09	0.23
3	11	0.50	0.73

Educational level (in years of education)	Absolute frequency (n)	Relative frequency (%)	Cumulative relative frequency (%)
4	100	4.55	5.28
5	156	7.09	12.37
6	169	7.69	20.05
7	221	10.05	30.10
8	450	20.46	50.57
9	251	11.41	61.98
10	320	14.55	76.53
11	479	21.78	98.32
12	31	1.41	99.73
13	6	0.27	100.00

Table 2 shows the distribution of educational levels among 18-year-old youngsters from Pelotas, Southern Brazil, with absolute, relative, and cumulative relative frequencies. In this case, absolute and relative frequencies correspond to the absolute number and the percentage of individuals according to their distribution for this variable, respectively, based on complete years of education. It should be

noticed that there are 450 adolescents with 8 years of education, which corresponds to 20.5% of the subjects. Tables may also present the cumulative relative frequency of the variable.

In this case, it was found that 50.6% of study subjects have up to 8 years of education. It is important to point that, although the same data were used, each form of presentation (absolute, relative or cumulative frequency) provides different information and may be used to understand frequency distribution from different perspectives.

When one wants to evaluate the frequency distribution of continuous variables using tables or graphs, it is necessary to transform the variable into categories, preferably creating categories with the same size (or the same amplitude). However, in addition to this general recommendation, other basic guidelines should be followed, such as: (1) subtracting the highest from the lowest value for the variable of interest; (2) dividing the result of this subtraction by the number of categories to be created (usually from three to ten); and (3) defining category intervals based on this last result.

For example, in order to categorize height (in meters) of a set of individuals, the first step is to identify the tallest and the shortest individual of the sample. Let us assume that the tallest individual is 1.85m tall and the shortest, 1.55m tall, with a difference of 0.3m between these values. The next step is to divide this difference by the number of categories to be created, e.g., five. Thus, 0.3m divided by five equals 0.06m, which means that categories will have exactly this range and will be numerically represented by the following range of values: 1st category - 1.55m to 1.60m; 2nd category - 1.61m to 1.66m; 3rd category - 1.67m to 1.72m; 4th category - 1.73m to 1.78m; 5th category - 1.79m to 1.85m.

Table 3 illustrates weight values at 18 years of age in kg (continuous numerical variable) obtained in a study with youngsters from Pelotas, Southern Brazil.[4,5] Figure 4 shows a histogram with the variable weight categorized into 20-kg intervals. Therefore, it is possible to observe that data from continuous numerical variables may be presented in tables or graphs.

TABLE 3

Weight distribution among 18-year-old young male sex (n = 2.194). Pelotas, Brazil, 2010

Weight at 18 years of age (in kg)	Absolute frequency(n)	Relative frequency (%)
40.5 to 59.9	554	25.25
60.0 to 65.8	543	24.75
65.9 to 74.6	551	25.11
74.7 to 147.8	546	24.89

Weight at 18 years of age (in kg)	Absolute frequency(n)	Relative frequency (%)

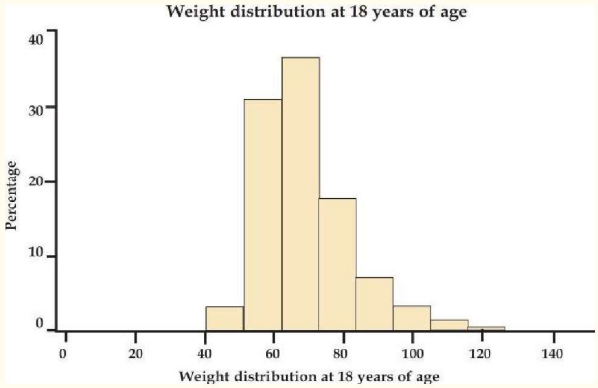


FIGURE 4

Weight distribution at 18 years of age among youngsters from the city of Pelotas. Pelotas (n = 2.194), Brazil, 2010

TABLE 4

Sun exposure during work and non-melanoma skin cancer (hypothetical data).

	Total		2.194		100.00

Work exposed to the sun	Non-melanoma skin cancer					Total
	Yes		No			
	N	%	N	%	N	%
20 or more years	30	75.0	10	25.0	40	100
<20 years	9	9.0	90	91.0	99	100
Never	1	0.3	300	99.7	301	100

Work exposed to the sun	Non-melanoma skin cancer					Total
	Yes		No			
	N	%	N	%	N	%
Total	40	9.0	400	91.0	440	100

The relationship between two numerical variables or between one numerical variable and one categorical variable may be assessed using a scatter diagram, also known as dispersion diagram. In this diagram, each pair of values is represented by a symbol or a dot, whose horizontal and vertical positions are determined by the value of the first and second variables, respectively. By convention, vertical and horizontal axes should correspond to outcome and exposure variables, respectively. Figure 5 shows the relationship between weight and height among 18-year-old youngsters from Pelotas, Southern Brazil, in 2010.[3,4] The diagram presented in figure 5 should be interpreted as follows: the increase in subjects'

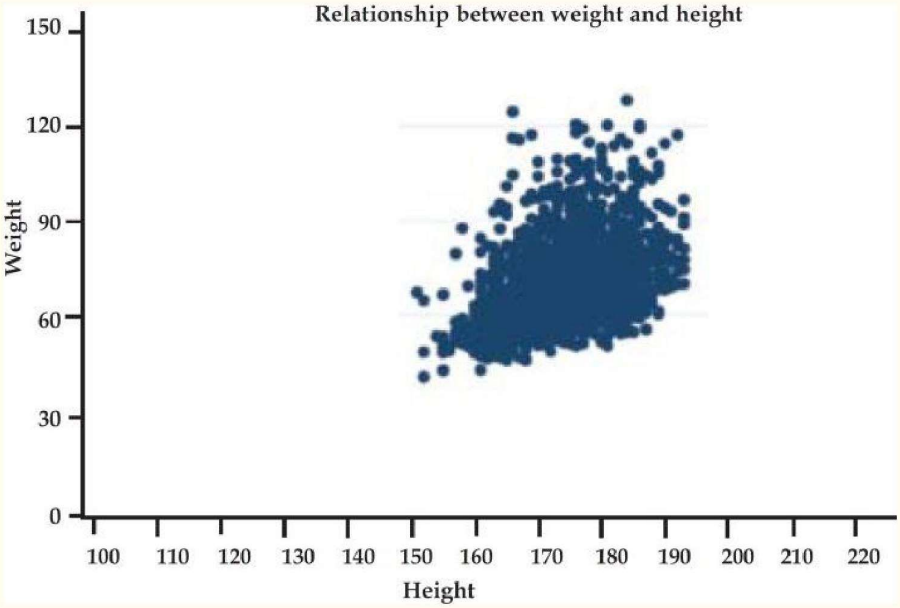


FIGURE 5

Point diagram for the relationship between weight (kg) and height (cm) among 18-year-old youngsters from the city of Pelotas (n = 2.194). Pelotas, Brazil, 2010.