DATA LITERACY

Getting Started

OFFICE OF PLANNING AND RESEARCH

Data Literacy I: Getting Started

In this module you will be introduced to foundational elements of data literacy. This training is the first in a 3-part series on data literacy. Each training series represents a building block in gaining the skills to read data, work with and explore data, and communicate about data.

Topics

01 The Importance of Data Literacy

02 The Role of Data Governance in Data Literacy

03 Data Types and Formats

04 Data Management and Storage

05 Data Sources

06 Data Ethics

07 Best Practices and Limitations

Learning Objectives

1. Describe what data is and the aspects of the world it represents
2. Identify the different sources of data available inside and outside your institution
3. Explain the ethical implications associated with being someone who uses data
4. Recognize that there are limits to data

01 | The Importance of Data Literacy

At Central Piedmont we aim to be a data literate organization

An institution that operates with data embedded in core decision making;

Where data is valued by leaders and colleagues alike; and

Where intentional and effective efforts to improve and sustain data literacy are prioritized with activities that build capacity.

We want to leverage that data and information to promote student success and improve alumni outcomes.

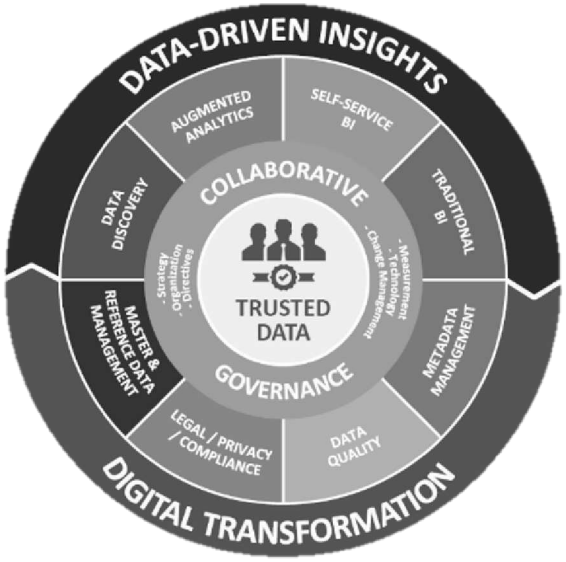
02 | Data Governance and Data Literacy

Data governance is an organization's processes and policies to gather, store, manage and dispose of data. A properly functioning governance structure ensures data privacy and enables collaboration across the organization.

Data literacy is the ability to read data, work with and explore data, and communicate about data by putting it in proper context. Data literacy includes understanding how to:

- ask the right questions
- evaluate data sources and collection methods
- interpret data to draw meaningful insights

Insight-driven decisions are supported by strong data governance and high levels of data literacy across an organization



03 | Data Types and Formats

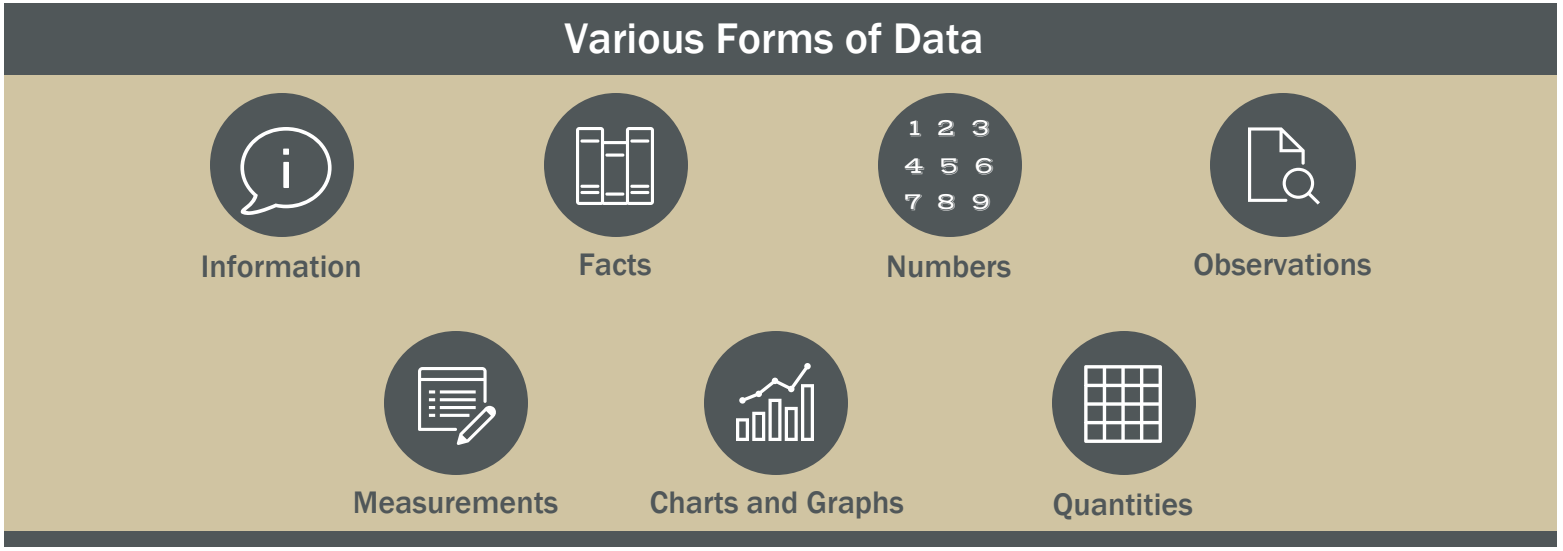
The process of turning data into meaningful insights begins with understanding the type of data you're working with.

Data can come in various forms and is classified into data types. This structure provides the framework for turning raw data into useful information.

03 | Data Types and Formats

What is Data?

Data is defined as a collection of individual facts or statistics, and can come in various forms such as numbers, observations, measurements, charts and graphs.



Data is a raw form of knowledge that becomes meaningful information when it is analyzed, organized and interpreted.

03 | Data Types and Formats

Types of Data

Data is classified into two general types: qualitative and quantitative. Identifying the types of data you're working with informs the nature and complexity of the insights you're able to glean from your data.

Qualitatative Data	Quantitative Data
Qualitative data is narrative in nature, describing qualities and characteristics.	Quantitative data deals with numbers or figures that can be counted or measured.
Provides answers to questions such as why? how?	Provides answers to questions such as what? when? who? how many? to what extent?
Gathers narrative information about participants' experiences and perceptions.	Uses percentages, averages (means) and other mathematical operations as a way of summarizing how a population thinks, feels, or acts.
The data can be collected using interviews, focus groups, open-ended survey questions.	The data can be collected through probability sampling, surveys, observation, document review.
Data is analyzed by categorizing and identifying themes and patterns.	Data is analyzed using statistical methods.
Subcategories: Nominal (labeled), Ordinal (ranked).	Subcategories: Discrete (finite integers), Continuous (any value within a range).

03 | Data Types and Formats

Data Type Subcategories

Quantitative and qualitative data are further classified into subcategories. Identifying the subcategory your data belongs to helps determine what kind of analysis can be carried out on the data.

Qualitative Data Subcategories

Nominal Data is used for labeling variables into distinct classifications; Doesn't involve a quantitative value or order.

Ordinal Data is used to depict the order or rank of non-mathematical ideas such as frequency, degree of satisfaction, education level.

Quantitative Data Subcategories

Discrete Data is a numerical type of data that refers to whole numbers that cannot be divided. In many cases, discrete data can be prefixed with "the number of."

Continuous Data is a numerical type of data that refers to numbers that can be broken into finer and finer units, usually within a range. Continuous data is about accuracy, with values often carrying decimal points.

Continuous data is further classified as Interval or Ratio data.

Interval Data is measured along a scale with familiar, constant, and computable differences; No true zero exists.

Ratio Data is similar to interval data, but ratio data has a true zero.

03 | Data Types and Formats

Summary of Data Types

Qualitative or Categorical Data

• Gender • Ethnicity • Marital Status • Zip Code	• Opinion (strongly disagree, disagree, neutral, agree, strongly agree) • Degree Level (Associate, Bachelor's, Master's, Doctorate) • Letter Grade (A,B,C,D,F)

Quantitative or Numerical Data

Discrete

Finite, numeric, non-negative integers; Usually expressed as "number of."

- Number of registered students
- Number of full-time employees

Continuous

Interval

Data measured along a scale with familiar, constant, and computable differences; No true zero exists.

- Calendar Years
- Age Range (<18,18-34, 35+)
- SAT Scores

Ratio

Similar to interval data, but ratio data has a true zero.

- Age
- Income
- Amount of time spent studying

03 | Data Types and Formats

Database Data Types

Database data types are attributes associated with a piece of data that tells a computer system how to interpret its value. Understanding data types ensures that data is collected in the preferred format and the value of each property is as expected--in other words, aiding clean and consistent data.

DATA TYPE	ATTRIBUTE	EXAMPLE
Integer (INT)	A numeric variable without a decimal. These are whole numbers.	0, 2, 3, 50
Float	All numbers including all integer values and every decimal and fraction of said integers.	7.8, 22/7
Character	An alphanumeric character that can be any letter, number or symbol.	'A' , 100
String	A sequence of characters from one character to multiple character values.	'Name'
Boolean	A binary data value.	True/False Yes/No
Date and Time	Date data type has: Year, Month, Day; can have the form YYYY-MM-DD. Time data type has the form HH:MM:SS (Hours, Minutes, and Seconds).	2022-09-28 10:07:33
Array	An ordered set of elements of a single data type. These can be created by the user or generated by an operation.	(1,2,3,4,5) or (a,b,c,d,e)

04 | Data Storage and Management

Effective data storage and management builds trust in the data being used to make decisions across an organization.

By ensuring that data is organized, reliable, accessible and secure, data management provides organizations with a means to easily evaluate important information in meaningful ways.

04 | Data Storage and Management

Types of File Permissions

One of the functions of data management is to set file permissions, which includes granting different levels of access to files. Files and directories can have three types of permissions:

Read allows a user to view the contents of a file or directory.

Write allows a user to change or modify the contents of a file. As it pertains to a directory, a user with write permissions will be able to add or delete files in a directory.

Execute allows a user to run an executable file, which is a file that contains an encoded sequence of instructions for a computer to perform.

04 | Data Storage and Management

Types of Data File Formats

A file format is a standard way that information can be encoded and stored on a computer or server, with some file formats requiring specific software to be handled.

TEXT

Text files are used to store plain text or tables of plain text.

Rich text format files can save more information than simple text. It can include formatting options and images.

CSV files are text files that can be used to store information as tables. Each row represents an observation and column values are separated by commas.

.txt .rtf .csv .html

DOCUMENT

Document files are like text files with the bonus that they can store more complex formatting information.

.doc .docx

SPREADSHEET

A spreadsheet file stores information in large tables (i.e., Microsoft Excel, Google Sheets). These tables can be used to perform calculations and conduct analyses.

.xls .xlsx .ods

OTHER

An xml file is a file used to structure data for storage and transport.

A pdf document is a general file format that cannot be modified but can still be easily shared and printed. Document processors are not needed to view or open these files; these can be viewed using a free viewer such as Acrobat Reader or through internet browsers.

.xml .pdf

04 | Data Storage and Management

Database Storage

A database is an organized repository of information that collects and stores data so that it can be easily accessed, managed and analyzed. Databases can be built for the sole use of an institution or for the general use of the public.

Private Databases

Private databases store information that is only available to specific individuals. Private databases are best suited for secured confidential information and core business systems.

Examples of data stored in a private database:

- Student registration and enrollment data
- Personnel records
- Financial information

Public Databases

Public databases store information that is available to the public for general use or analysis. Government organizations are usually stewards of public databases.

Examples of information available from public databases:

- Enrollment trends at other institutions
- Labor market trends
- Population statistics and demographic information segmented by state, city and even zip code

04 | Data Storage and Management

Types of Databases

Database structure can vary, so it's important to know the different types.

A **centralized database** is located and maintained in a single location. This location is usually a central computer, or server CPU and is typically utilized by businesses or higher education institutions.

Unlike centralized databases, a **distributed database** is stored across multiple locations, though it may appear to be a single database to users. These stores are connected via communication links, allowing users to access the data easily across platforms.

A **cloud database** is stored in a virtual environment and commands are executed over the cloud computing platform. Some examples of a cloud computing platforms include Amazon Web Services, Google Cloud SQL, Microsoft Azure, etc.

In a **relational database**, information is stored in tabular format (rows and columns) with keys (set of attributes) being used to establish relationships between different tables. This system uses SQL (Structured Query Language) to store, manipulate and maintain the data.

A **network database** houses information using a graph structure where data is in the form of nodes connected by vertices (links) between them. This is similar to a Hierarchical database, but each data point can have multiple children and parent nodes.

A **document database** is a type of nonrelational database that is used to store and query documents that are encoded using a standard format like XML or JSON.

04 | Data Storage and Management

Data Warehouse

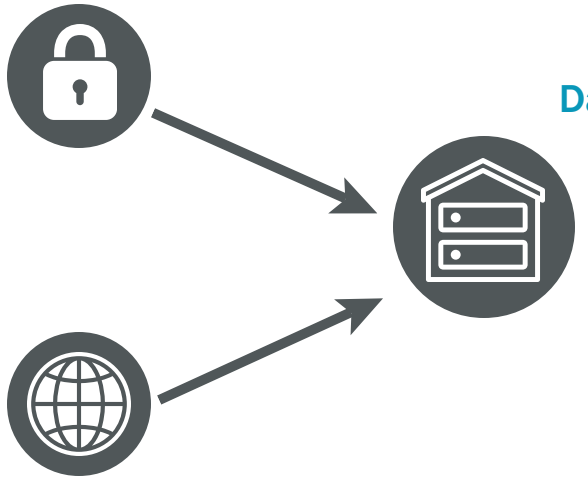
A data warehouse is a centralized repository of digitally stored business information used to drive reporting and data analysis. It consists of massive quantities of data extracted from numerous sources, such as relational databases, internal applications, transactional systems, and external partner systems, aggregated into unified, well-structured historical data.

Private Databases

- Student Data
- Personnel Data
- Financial Data
- Marketing Data

Public Databases

- Competitor Information
- Industry Information
- Labor Market Data
- Population Information



Data Warehouse

- Student Data
- Personnel Data
- Financial Data
- Marketing Data
- Competitor Information
- Industry Information
- Labor Market Data
- Population Information

05 | Data Sources

Information from multiple sources brings you closer to the truth.

Data sources can include data that are already collected and data that will be collected during the research process.

By looking at a diversity of data you build more diverse thinking, and diverse thinking leads to more effective decision making.

05 | Data Sources

Internal and External Databases

Organizations gather and collect data from many sources, both inside and outside of their own institution. There are two main classifications of these datasets.

Internal Databases

An internal database is any collection of internal data that is generated within one's institution.

Central Piedmont uses many internal tools to collect and compile information regarding student and employee records, business and community connections, and billing information.

External Databases

An external database is any external collection of information generated for specific uses and situations made outside one's own organization.

Some information collected includes data on business trends and hiring practices, productivity and labor trends, and population healthcare dynamics and trends.

05 | Data Sources

Internal Data Sources

Colleague
The college's Student Information System (SIS); Colleague is the single source of truth for Central Piedmont data.

Informer
Tool used to extract and normalize data from Colleague, the college's student information system.

Ad Astra
Course scheduler tool used by Central Piedmont to manage space and faculty resources (i.e., room reservations).

Watermark Student Success and Engagement
Tool used by Central Piedmont to keep underserved students engaged, increase retention rates and optimize degree and certificate completion.

Brightspace
Central Piedmont's online Learning Management System (LMS). Brightspace replaced Blackboard as the college's LMS in 2021.

NOTE: Login is required for Central Piedmont internal data sources. Access approval may also be required. Contact the Service Desk for assistance.

External Data Sources

North Carolina Community College System Dashboard
The North Carolina Community College System (NCCCS) interactive dashboards allow users to explore and interact with NCCCS data by theme. Includes system and college-level data views, disaggregations, historical trends, and peer comparisons.

University of North Carolina Systems Office Dashboard
Interactive reports that provide the public with access to detailed system data on selected core measures (enrollment trends across the UNC system, trends in transfer student enrollment, Freshmen admissions and performance, degrees awarded, educator quality).

National Center for Education
The primary federal entity for collecting and analyzing data related to education in the U.S. and other nations. The NCES collects, analyzes and reports on complete statistics on the condition of U.S. education, as well as education activities internationally.

U.S. Bureau of Labor Statistics
Measures labor market activity, working conditions, price changes, and productivity in the U.S. economy to support public and private decision making.

U.S. Census Bureau
Provides data on the nation’s people and economy. Includes population data by demographic breakdown, business data by industry, geography and more.

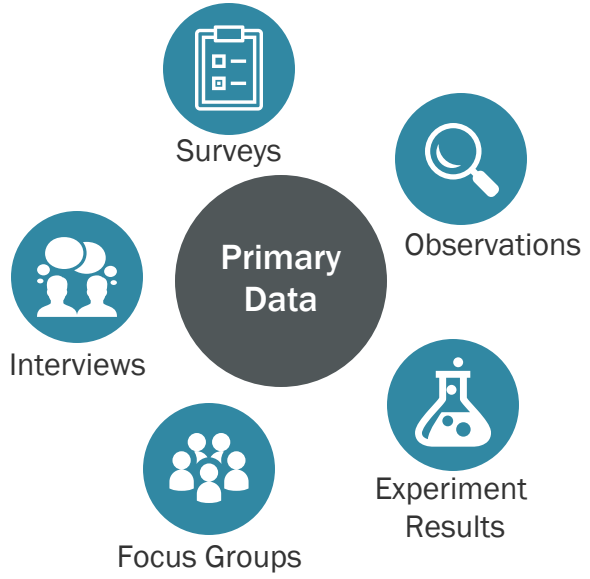
05 | Data Sources

Primary and Secondary Data

Primary data and secondary data both have their advantages and disadvantages. Knowing the methods or sources used for these types of data can help you make proper decisions when choosing primary or secondary data for your research work.

Primary Data

From the perspective of a researcher, primary data refers to **any firsthand data generated by the researcher themselves**, using methods like experimentation, surveys or observations.



Secondary Data

From the perspective of the researcher, secondary data means **any data collected by someone else earlier.** Some examples can be data gathered from different public databases.

05 | Data Sources

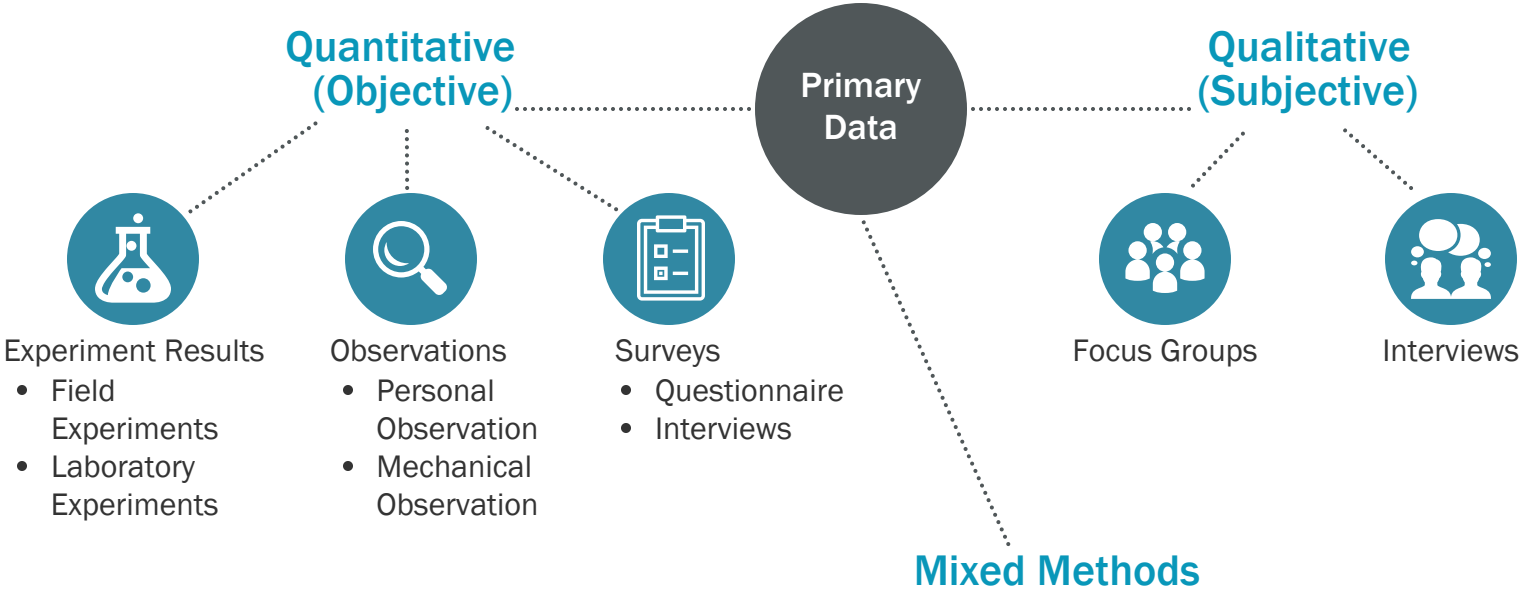
Primary Data: Quantitative and Qualitative Data Sources

Quantitative Data Sources

Quantitative data sources include specific measurements and outcomes collected from surveys, observations, experiment results and other numeric measurements. **These data sources can be categorized as objective** data sources.

Qualitative Data Sources

Qualitative data sources include information collected from interviews and focus groups. **These data sources can be categorized as subjective** data sources and focus on the subject's perceptions and experiences.



05 | Data Sources

Secondary Data: Considerations

Precautions that must be considered when handling secondary data

- Is the data **relevant** to the research on hand?
- Is the data from a **reputable** source?
- Is the data **accurate**?
- **How was the data collected** and has the process been documented appropriately (e.g., when reviewing results of an experiment)?
- What was the **objective of the original study**?
- Do **clusters or stratifications** need to be accounted for before implementing analyses?
- **When was the data collected** and what was the condition of the subjects?

06 | Data Ethics

Good intentions are not enough to use data ethically.

- Jeffrey Alan Johnson
Association of Institutional Research

Data ethics is a branch of ethics that evaluates data practices—collecting, generating, analyzing and disseminating data—that have the potential to adversely impact people and society.

Increased awareness of data ethics enable organizations to operate in an ethical manner that abides by regulations, builds trust with users, and keeps personal data private.

06 | Data Ethics

Twelve Principles of Data Ethics

The highest priority is to respect the persons behind the data.

Always follow the law, but understand that the law is often a minimum bar.

Data scientists and practitioners should accurately represent their qualifications, limits to their expertise, adhere to professional standards, and strive for peer accountability.

Attend to the down-stream uses of datasets.

Be wary of collecting data just for the sake of more data.

Aspire to design practices that incorporate transparency, configurability, accountability, and auditability.

Provenance of the data and analytical tools shapes the consequences of their use.

Data can be a tool of inclusion and exclusion.

Products and research practices should be subject to internal, and potentially external ethical review.

Stive to match privacy and security safeguards with privacy and security expectations.

As much as possible, explain methods for analysis and marketing to data disclosers.

Governance practices should be robust, known to all team members and reviewed regularly.

Source: https://www.ethicalresolve.com/blog/12principlesofdataethics.html

06 | Data Ethics

Summary of Transparency and Intention

Trust

Data transparency helps facilitate proper interpretation and dissemination of results by other stakeholders. The practice of data transparency includes disclosing how information is collected and the purpose of that collection, as well as the process used to analyze and communicate that information.

The intention of a data collection and statistical study should always be made clear, in the documentation and in the communication. When we speak of intention, we should always ask ourselves “why?” for everything.

- Why does this particular data answer a question?
- What do the authors hope to communicate to the reader using this data?
- Why should I care?
- Why would anyone care?

06 | Data Ethics

Summary of Privacy and Security

Compliance

Data privacy defines who has access to particular types of data and regulations that ensure companies and institutions take measures to protect private user data.

The U.S. government has passed laws governing the privacy of individual's data. One example is the Family Educational Rights and Privacy Act (FERPA), which protects the privacy rights of students. Another example is The Health Insurance Portability and Accountability Act of 1996 (HIPAA) which protects sensitive patient health information from being disclosed without the patient's consent or knowledge.

Data security involves of set of processes and procedures used in safeguarding information throughout its life cycle. Safeguarding includes protections from corruption, theft and unauthorized access.

06 | Data Ethics

Summary of Data Quality and Unintended Human Biases

Fair Practices

Always make sure that the data is appropriate for the question. Also, make sure the data is as representative and accurate to the population --and as complete-- as possible.

Representative and accurate means ensuring that biases and errors do not compromise the potential benefits of a project or mislead data users, which further necessitates that the appropriate data be collected in order to answer specific questions.

Extrapolation from limited data is problematic from a researcher standpoint and policy standpoint. If information is biased or incomplete, this may lead to uninformed measures and policies.

Disaggregating data by subpopulations of society might give more insight into particular groups of individuals and allow for a more robust statistical analysis. However, it is important to remember that putting any sub-population under a microscope can raise ethical concerns. The results of such studies should be shared with caution and always aligned with the necessary context and intention.

07 | Best Practices and Limitations

The more high-quality data you have, the more confidence you can have in your decisions.

By following best practices and being aware of the limitations of data, you can take actions to ensure the use of reliable data while identifying ways to overcome limitations.

Good data decreases risk and generates actionable insights.

07 | Best Practices and Limitations

Statistics: Ten Simple Rules for Effective Statistical Practice

Statistical knowledge helps you use the proper methods to collect data, employ the correct analyses, and effectively present the results. Simple rules for best practices in statistics are summarized below. Read the full article by Kass et. al., from Carnegie Mellon for details about each rule.

Statistical methods should enable data to answer scientific questions. Start with the underlying research question and from there consider the ways data might provide answers.

Signals always come with noise. Capturing the variability in data (noise) can call attention to likely sources of systematic error, known as bias.

Plan ahead, really ahead. Ask questions at the design stage: careful data collection can greatly simplify analysis, and make it more rigorous.

Worry about data quality. Understand your data: why are some data missing or incomplete; how was the data selected prior to formal analysis, and how might such selection affect conclusions?

Statistical analysis is more than a set of computations. Explain your choice of statistical method in answering your research question.

Keep it Simple. Good design, implemented well, can often allow simple methods of analysis to produce strong results.

Provide Assessments of Variability. When reporting results, it is essential to supply some notion of statistical uncertainty.

Check your assumptions. A variety of factors can cause violations of statistical modeling assumptions, even in the best analyses.

When possible, replicate! The scientific results that stand the test of time are those that get confirmed across a variety of different, but closely related situations.

Make your analysis reproducible. Improve the ability to reproduce findings by being systematic about the steps in the analysis and sharing the data and code used to produce the results.

07 | Best Practices and Limitations

Data Management: Seven Best Practices for Successful Data Management

Data management is a critical process to ensure data is acquired, validated, stored and protected. Best practices in data management are summarized below. Read the full article by data specialists at Tableau Software, LLC for details about each best practice.

Build strong file naming and cataloging conventions. If you are going to utilize data, you have to be able to find it.

Carefully consider metadata for data sets. Metadata is descriptive information about the data you are using. It should contain information about the data's content, structure, and permissions so it is discoverable for future use.

Data Storage. If you ever intend to be able to access the data you are creating, storage plans are an essential piece of your process. A solution that works for a huge enterprise might not be appropriate for a small project's needs, so think critically about your requirements.

Documentation. Produce multiple levels of documentation that will provide full context to why the data exists and how it can be utilized.

Commitment to data culture. Make sure that data literacy is prioritized throughout the organization.

Data quality trust in security and privacy. Building a culture committed to data quality means a commitment to making a secure environment with strong privacy standards.

Invest in quality data-management software. Using a reliable software that helps you build, catalog, and govern your data will build trust in the quality of your data.

07 | Best Practices and Limitations

Statistics Limitations

It isn't an "exact" science

Statistics is built on probability, meaning that conclusions that are draw usually follow the central tendency of the data. Though this is fine, it's important to realize that this value is not exact.

Studies the group and not the the individual

Statistics are based on aggregate facts and data; a single numerical measurement cannot be called statistics implying that the study deals with groups and not the individual.

Statistics are true on average

Similar to the above statement, statistics are true on average. In other words, because of the variability between samples and subjects, comparisons are made by averages, even though these averages may be composed of individual data points significantly different from each other.

Statistics cannot be performed on heterogeneous data

Statistics cannot be applied to heterogeneous data because it cannot be compared to other data sets. The similar features of data is the mechanism which facilitates comparison and analysis among different observations.

Statistics is better applied to quantitative data

Statistical methods are best applied to quantitative data because it is difficult to quantify qualitative characteristics using a common system. It can be applied to qualitative data in some instances but extra precautions should be taken before results are disseminated.

07 | Best Practices and Limitations

Data Limitations

Incomplete or missing data

Incomplete, or missing, data can severely limit its usability and can potentially lead to biased estimates or invalid conclusions.

Self Reporting

Self-Reporting: if you are using data collected from surveys and other self reporting tools, keep in mind that people don't always provide accurate information.

Data quality varies

Data can vary in quality and format. Data collected from diverse sources (different databases, emails, surveys, etc.) will have different attributes and structures and may not be compatible with institution defined data fields.

Uncertainty is inevitable

No matter how much data you have, you can never be 100% sure of anything so be careful of drawing definitive conclusions.

Context is everything

Context is not only critical for data interpretation and visualization, but critical in data collection and analysis.

Data can tell a lot of stories

Data can tell lots of different stories, but it is important to remember that data alone tells you nothing. It is only through analysis you can uncover insights.

07 | Best Practices and Limitations

Correlation ≠ Causation

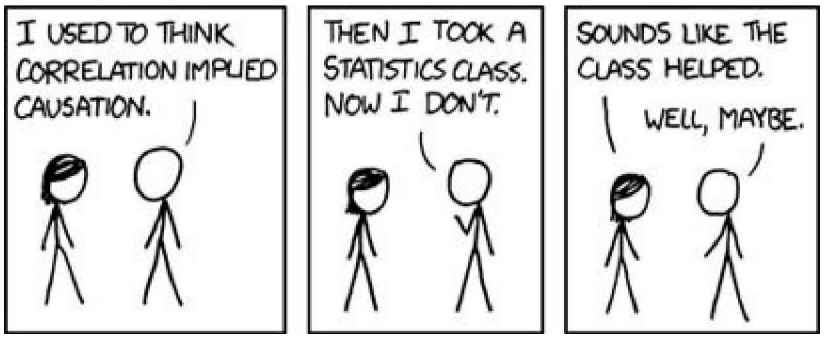
Correlation is a measure describing the size and direction of a relationship between two variables.

Causation means that one event is the result of the occurrence of some other event (cause and effect).

As tempting as it may be to say otherwise, correlation does not imply causation.

Common Pitfall

"Correlation proves causation" is a logical fallacy known as *cum hoc ergo propter hoc*, Latin for "with this, therefore because of this." In other words, assuming that because two events happened at the same time, then one event caused the other event to happen. This is a common pitfall that can occur when a cause and effect relationship is drawn between events happening at the some time, without investigating possible alternate explanations for the correlation.



07 | Best Practices and Limitations

Correlation does not imply Causation!

A strong correlation between two variables does not necessarily imply a causal relationship between them. Consider the following examples:

A Drowning deaths and ice-cream sales are strongly correlated, as ice cream sales increase then so do drowning deaths. Does this mean that the sale of ice cream causes drownings? No!

The correlation comes from the fact that both are affected by the season or time of year.

B In the 1990's there were numerous epidemiological studies that showed women who were taking a combined hormone replacement therapy had a lower-than-average chance of coronary heart disease. Even though coronary heart disease is associated with socioeconomic status, these studies led doctors to propose hormone replacement therapy was protective against coronary heart disease.

Randomized controlled trials showed that HRT caused a statistically significant increase (though small) in risk of coronary heart disease. Further investigation showed that women who could afford to take HRT were from a higher socioeconomic class with healthier dietary habits and exercise regiments. HRT and decreased risk of coronary heart disease were coincident effects rather than cause and effect.

Article — Commentary: The hormone replacement–coronary heart disease conundrum: is this the death of observational epidemiology https://doi.org/10.1093/ije/dyh124

07 | Best Practices and Limitations

If Correlation doesn't imply Causation, what could it be?

To avoid committing the *cum hoc ergo propter hoc* fallacy, rule out other possible explanations for the correlation.

Causal link
Even if there is a causal link, correlation does not tell you x causes y (think back to definition of correlation).

Hidden cause
A hidden variable can cause both x and y, creating the correlation (think back to ice cream and death example).

Confounding variable
Another variable z can affect y with x. So, results depend on the confounding variable as well (think about the hormone replacement therapy and coronary heart disease example).

Coincidence
Correlation that occurred by chance.

Test Your Understanding

Test Your Understanding

Data Literacy Fundamentals

1 **Which of the following is an example of an ordinal data type?**

a) Student Age
b) Letter Grade
c) Major of Study
d) Enrollment Status

2 **Which of the following is not an internal database used at Central Piedmont?**

a) Watermark
b) Colleague
c) NCES
d) Brightspace

3 **What law passed by the federal government protects the privacy of students?**

a) Individuals with Disabilities Education Act
b) Family Educational Rights and Privacy Act
c) Higher Education Act
d) Elementary and Secondary Education Act of 1965

Test Your Understanding

Data Literacy Fundamentals

4 **Choose the correct answer for the following statement:**
Correlation implies causation

a) True
b) False
c) Not enough information

5 **Label the following variables as qualitative or quantitative:**

a) Student age
b) Student ID number
c) Marital Status
d) Number of siblings
e) Phone number

6 **Consider the following scenario:**
You created a file for your department containing cleaned data from a long-term project and you want to share this information to others within the department. Because this took a lot of work you only want others in the department to be able to view the contents of the file, but not change anything in the file.

What file permissions should you grant to the other members of your department?

a) Read and write privileges
b) Read only privileges
c) Read, write, execute privileges
d) Write only privileges

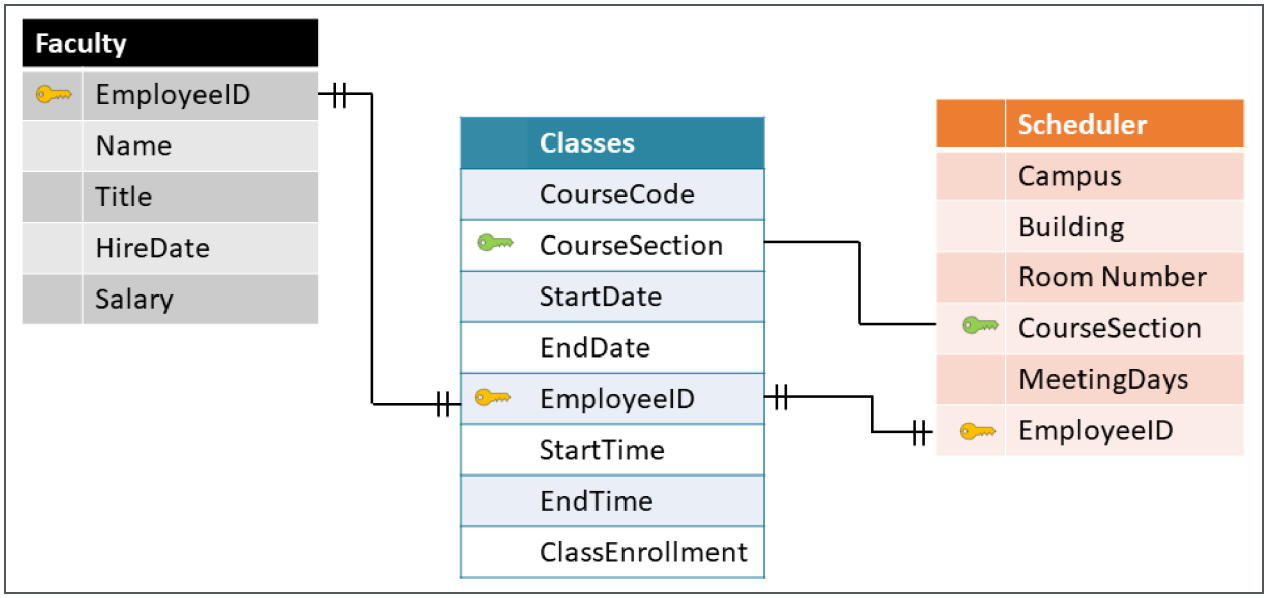
Test Your Understanding

Data Literacy Fundamentals

7 **Consider the database structure below. You notice all the tables in the database connect to each other using the common variable EmployeeID.**

What type of database is this?

a) Cloud database
b) Network database
c) Relational database
d) None of the above



Test Your Understanding

Data Literacy Fundamentals

8 **Consider the following hypothesis and research study:**

People that get married and live together are more likely to gain weight than those who are single and live alone. In order to determine if this is a true statement, researchers observed 5,000 individuals for 5 years. At the start of the study, everyone was single. Researchers found that people who were partnered and living together gained on average 10 more pounds than people that remained single and lived alone.

Based on the result of the study, can we conclude that getting married and cohabitating causes weight gain?

a) Yes, the results of the study have established a cause-and-effect relationship between getting married and weight gain.

b) Yes, though the observational study has not established a cause-and-effect relationship between getting married and weight gain, it is obvious that getting married causes weight gain.

c) No, you cannot conclude a cause-and-effect relationship between marriage and weight gain using an observational study.

d) No, even though we know that there is a cause-and-effect relationship between marriage and weight gain based on the results of the study, if we said this out loud, we would get in trouble.

Test Your Understanding

Data Literacy Fundamentals

9 **A record, document, diary, manuscript, or other source of information created by the researcher at the time of study is considered:**

a) Primary Data
b) Secondary Data

10 **A delimiter is a character that marks the end or beginning for a unit of data. Which of the following file formats use a comma as a delimiter?**

a) .docx
b) .csv
c) .txt
d) .xlsx

Next Steps

Congratulations on completing Data Literacy I: Getting Started. You have gained important foundational knowledge in data literacy and are ready to apply that knowledge to the next steps. Like any skill, data literacy involves multiple levels of proficiency. You can further develop your data skills through exposure, study and practice.

Data Literacy Training and Workshops

The Office of Planning and Research developed a 3-part Data Literacy training series. Each module in the series represents a building block for gaining data literacy skills. Planning and Research also hosts an annual data summit with workshops focused on current data topics and trends.

- Sign up for additional data literacy training modules:
    - Data Literacy II
    - Data Literacy III
- Attend the annual **Data Summit**

Online Resources

Access to dashboards, reports, glossary of terms, internal and external data sources, and other information about institutional research and institutional effectiveness is available on the Office of Planning and Research website and intranet.

Contact Planning and Research

Questions? Contact the Office of Planning Research at research@cpcc.edu.


CENTRAL PIEDMONT
COMMUNITY COLLEGE