Unit 2- LITERATURE SURVEY AND DATA COLLECTION

SYLLABUS:

Importance of literature survey - Sources of information - Assessment of quality of journals and articles - Information through internet. Effective literature studies, approaches, analysis, plagiarism, and research ethics. Data - Preparing, Exploring, examining and displaying. 5 Hours

2.1 IMPORTANCE OF LITERATURE SURVEY

(Q: What is the significance of conducting a literature survey after defining a research problem?)

Once the problem is clearly defined, the researcher should undertake an extensive survey of the background literature related to the relevant theories, articles, reports, records, etc. This is done to narrow down the problem and to find the gap in the literature, to find out what material and other data are available for operational purposes, the problem and difficulties faced by previous researchers in their study etc. All these will enable a researcher to move up starting from the existing premise, sharpen his focus of attention on specific aspects, develop new ideas, formulate general approach to the given problem, techniques that might be used, possible solutions etc.

2.1.1 Sources of Information

(Q: Identify the sources of information that are helpful for conducting literature survey and classify them.)

- The sources of information are classified as: 1. Public sources: i. Central Government Departments (Defense, Energy, Science and Technology, etc.) ii. State and local Government (Highways, Pollution Control Board, etc.) iii. Libraries iv. Universities v. Internet.
  2. Private sources i. Nonprofit organizations (professional societies, trade associations) ii. Profit-oriented organizations (manufacturers, consultants, vendors' catalogues, samples, test data, etc.,) iii. Individuals (friends, faculty).
- The major sources of information are the library and internet. They provide: • Technical dictionaries • Encyclopedias • Handbooks • Bibliographies • Indexing and abstract services • Technical and professional journals • Translations • Technical reports • Books • Patents • Catalogues and manufacturer's brochures.
- Libraries and research institutes provide indexing and abstracting services to retrieve published literature. An "indexing service" cites the article by title, author, and bibliographic data. An "abstracting service" provides a summary of the content of the article, Ex: Ceramic Abstracts, Fuels and Energy Abstracts, Highway Research Abstracts etc. "Dissertation abstract" gives abstracts of doctoral dissertations compiled.
- The various sources from which information is gathered are grouped into (i) Primary sources: These provide first-hand or original information. Ex: data gathered out of an experimental study, survey data, interview transcript etc. ii) Secondary sources: These contain information pulled out from primary sources and these include abstracts, review articles, etc. (iii) Tertiary sources: The information provided by them are taken from primary and secondary sources, Ex: textbooks, handbooks, etc.

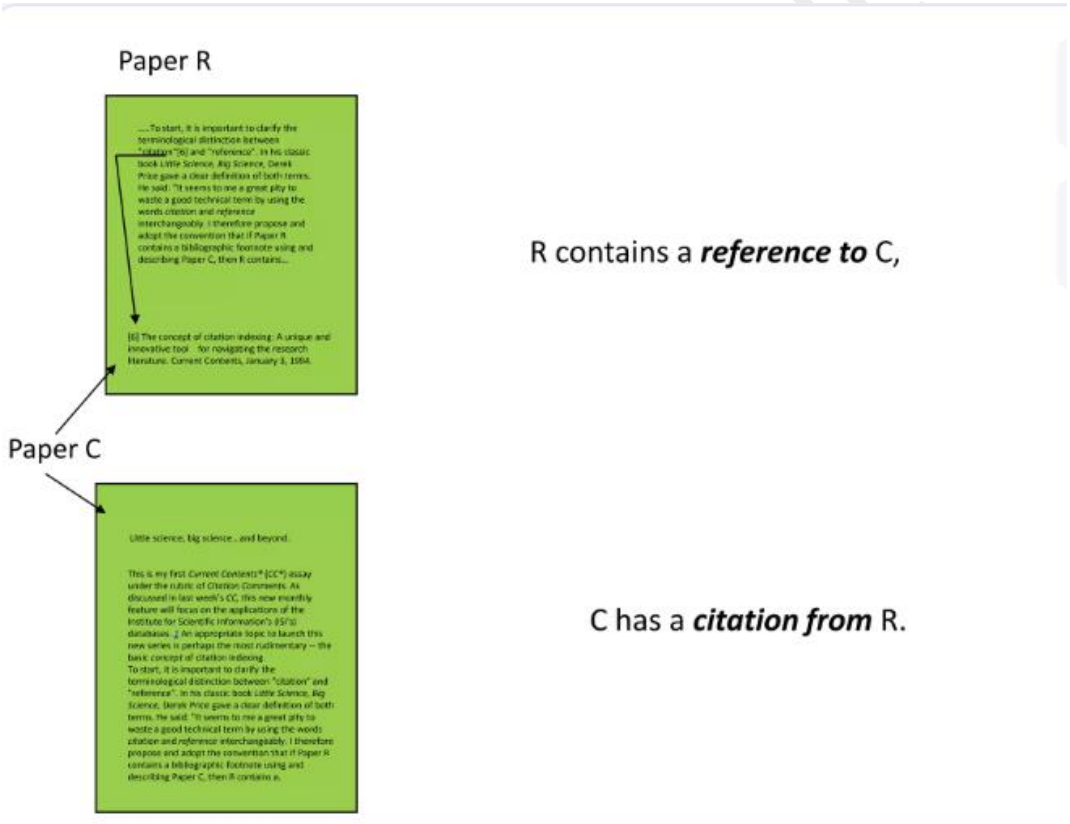
2.1.2 Databases and Citation Indexes for Research

(Q. Give the meaning of 'citation', 'reference' and 'citation index')

(Q: How can researchers effectively utilize computer-aided databases and citation indexes to retrieve relevant information and identify key documents in their field of study?)

2.1.2.1 Concepts of 'citation', 'reference' and 'citation index' and difference between them

- **Citation in research papers**: A **citation** appears in the main text of the paper. It is a way of giving credit to the information that is specifically mentioned in research paper by leading the reader to the original source of information. The researcher needs to use citation in research papers whenever there is a need to elaborate a particular concept in the paper, either in the introduction or discussion sections or as a way to support research findings in the results section. The references cited in a paper are sources of information that support the paper's arguments, assertions, or quoted text. By citing the references in the scientific papers, authors make explicit linkages between their current research and prior work present in the scientific literature.
- **Reference in research papers**: A **reference** is a detailed description of the source of information. The references in research papers are in the form of a list at the end of the paper. The essential difference between citations and references is that citations lead a reader to the source of information, while references provide the reader with detailed information regarding that particular source.



adopted from : Mathew, N. (n.d.). Citation indexing. Retrieved from http://ist.psu.edu/faculty_pages/giles/IST497/presentations/M

Figure 2.1 Concept of citation considering papers 'R' and 'C'

Ex: If a paper R as shown in figure 2.1 uses an information found previously in paper C published prior to paper R, then:

- ✓ R contains a reference to C

✓ C has a citation from R

R (published in 2020) is citing article and C is cited article (published in 2018(should be < 2020))

- **Bibliography in research paper**: It is a list of sources that appears at the end of a research paper, and contains information that may or may not be directly mentioned in the research paper. The difference between references and bibliography in research is that an individual source in the list of references can be linked to an in-text citation, while an individual source in the bibliography may not necessarily be linked to an in-text citation.
- **A citation count** is the frequency of an article cited by other articles Ex: If a paper 'C' is cited in R1, R2…R5, then citation count of C =5. Figure 2.2 graphically shows how citations work.

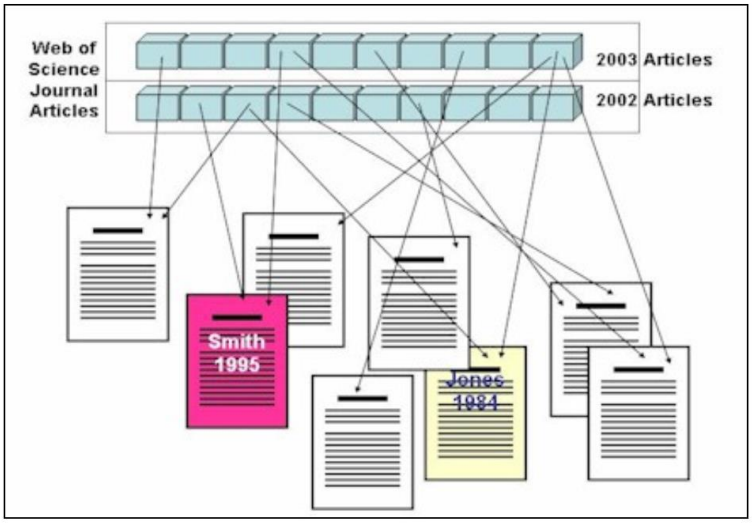


Figure 2.2 Concept of citation

A citation index is a bibliographic database connecting citing articles to cited articles. It allows the user to easily establish which later documents cite which earlier documents. While in a given paper its reference list points to earlier work as influences, only a citation index can provide a list of the later papers that cited the given paper.

(An **index** is an alphabetical listing of the topics in a document, and is located at the end of a document)

2.1.2.2 The use of citation index:

Citation indexes are used for citation analysis.

For deciding the quality of paper, journal: The citation index provides a useful criterion for judging the quality of a paper published in a journal. Citations are used:

- as a measure of importance of the information source
- to gather data on the "impact" of journals
- for assessing and analyzing particular areas of research activity and publication.
- for evaluating a journal's impact factor
- for deciding the author's professional standing

Missing References using "Citation Index": Once the researcher finds a reference of interest, he can use citation index to find all other references that cited the key reference.

For identifying the key documents: Identify the references which have the greatest number of citations, or those that other experts in the field cite a particularly important. It takes 6 to 12 months for a reference

to be included in an abstract service, so current research will not be picked up using this strategy. Current research can be obtained by searching using keywords, subject heading, journal titles and authors already identified from literature search.

2.1.3 Assessment of Quality of Journals and Articles

(Q: How do the impact factor and h-index serve as indicators of journal and article quality in the context of research methodology?)

Impact factor of the journal: The quality of a journal is judged by the impact factor of the journal. The impact factor of a journal is calculated by dividing the references cited in one year by the number of citable articles published in the same journal over the previous two years.

Ex: 2020 Impact Factor (IF) Calculation:

Year for which IF to be calculated = 2020

Previous two years to be considered =2018, 2019

No. of articles published in that journal during 2019 =311

Total no. of articles published in that journal during 2018 =356

Total no. of citations for the 356 articles published during 2018 =3485

Total no. of citations for the 311 articles published during 2019 =2516

Then Journal IF (JIF) is given by:

$$\text{Journal IF for 2020} = \frac{\text{Total number of citations in 2020 for the articles published in 2018+2019}}{\text{Total number of articles published in 2018+2019}}$$

$$= \frac{3485+2516}{356+311} = \frac{6001}{667} = \mathbf{9}$$ (Adapted from https://clarivate.libguides.com/jcr)

This ratio is published annually in Thompson Scientific Journal Citation Reports (JCR). The scientific JCR impact factors are based on data from Journals indexed in Web of Science. Scopus and Web of Science are global databases for **discovering academic literature.** They allow users to s**earch for articles, conference proceedings, trade publications, and book chapters on a topic**. They also help to find author information, such as 'h'-index, and lists of publications.

The calculation of 'h'-index: SCOPUS uses a measure called "h-index", which was developed by Hirsch in 2005 to evaluate the impact of journals. h-index is based on a list of publications ranked in descending order by the citation count. The value of h is equal to the number of papers (N) in the list that have N or more citations. One can obtain the h-index using freely available databases-Google Scholar and Publish or Perish, subscriber databases such as SCOPUS and Web of Science. The h-index can also be used to evaluate the productivity and impact of a group of scientists, in a Department or University or a country. Science Citation Index (SCI) is owned by Thompson Reuters and is called "Science Citation Index Expanded".

It covers more than 6500 world's leading and significant journals across more than 150 disciplines and the selection process is very rigorous. It is available through Web of Science database, a part of Web of Knowledge collection databases.

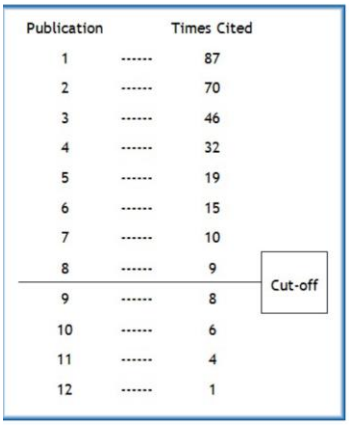


Figure 2.3 h-index calculation

Example for calculation of h-index:
As shown in figure 2.3, the publications of a researcher are arranged in decreasing order to determine the h-index. As per that the researcher has an h-index of 8, as 8 articles have been cited at least 8 or more times.

2.1.4 Information through Internet

(Q: How researchers can ensure comprehensive information retrieval through the World Wide Web and various search engines?)

The internet is a network of connected local computer networks using common protocols. A key part of it is the World Wide Web (www), which consists of linked documents accessed through hypertext. Web browsers like Microsoft Internet Explorer, Google Chrome, Mozilla Firefox, Apple Safari and Microsoft Edge allow users to navigate the web.
Web addresses are identified by Universal Resource Locators (URLs). They identify specific files on the internet. Ex: URL http://www.isoc.org/internet-history points to a file called "internet-history" on the isoc.org server, with the prefix http://www indicating an HTTP server on the World Wide. To find information on the web, we use search engines, which sort and index data differently, so results may vary. A study found that even the best search engine covers only about one-third of all web pages, so using two or three search engines is the best way to maximize information retrieval.
The most commonly used search engines include,
• Alta Vista http://www.altavista.digital.com
• Google Scholar http://scholar.google.com
• Microsoft Academic http://academic.research.microsoft.com
Search engines like Scirus, SciNet, Scholar provide for most comprehensive scientific and technology information.

The researcher should also have knowledge of many URLs devoted to engineering topics. Some examples of URLs are, NASA Technical Information Service http://tech reports.lors.nasa.gov/egi-bin/ NTRs. National Technical Information Service http://ntis.gov. Because of the vastness of open literature and the inaccessibility of many proprietary files, one can never be sure that all works have been uncovered. Therefore, a sincere effort in doing a thorough literature survey is essential.

2.2 EFFECTIVE LITERATURE STUDIES, APPROACHES and ANALYSIS

2.2.1 Literature study (review):

(Q: What is a literature review? Outline the different steps involved in the process?)

A literature review is a comprehensive review and analysis of the published literature on a specific topic or research question. The literature that is reviewed contains: books, articles, academic articles, conference proceedings, association papers, and dissertations. It contains the most pertinent studies and points to important past and current research and practices. It provides background and context, and shows how your research will contribute to the field.

2.2.1.1 Steps in a literature study:

1. **Developing a research question/objective in a specific subject area**, that is:
   - Focused
   - Not too broad and not too narrow in scope

- Complex enough to allow for research and analysis

2. **Searching the existing literature:** It refers to searching the literature and making decisions about the suitability of material to be considered in the review. (There are three coverage strategies: 1. Exhaustive coverage: An effort is made to be as comprehensive as possible in order to ensure that all relevant studies, published and unpublished, are included in the review and, thus, conclusions are based on this all-inclusive knowledge base. 2. This coverage searches for relevant articles in a small number of top-tier journals in a field. 3. This strategy, the review team concentrates on prior works that have been central or pivotal to a particular topic.). It includes following steps:

i. Make a list of relevant keywords and phrases: Keywords are **words used to search the record of an article, book, or other material in library databases**. To identify keywords:

1. Write down the research statement or question.
2. Underline the two or three most important terms that represent the topic
3. Use synonyms of initial keywords as additional search terms.

Ex: Are social media users concerned about their personal privacy?

ii. Searching with keywords

- Use **AND** to link *different* keywords together
- Use **OR** to group synonyms, *similar* concepts together in parentheses
- Use **quotation marks** to search for *specific phrases*, or key words with *two or more words*

Ex: **("Social Media" OR "social network") AND (privacy OR "personal privacy") AND (concern OR worry)**

3. Screening for inclusion: The identified material must be screened for relevance using a set of predetermined rules that determine which studies are included or excluded

4. Assessing the quality of primary studies: Researcher should assess the scientific quality of the selected studies to refine which studies to include in the final sample.

5. Extracting data: This involves extracting applicable information from each primary study included in the sample and deciding what is relevant to the problem of interest.

6. Analyzing and synthesizing data: Researcher must collect, summarize, aggregate, organize, and compare the evidence extracted from the included studies. The extracted data must be presented in a meaningful way that suggests a new contribution to the existing literature.

2.2.1.2 Types of Review

(Q: What are the various kinds of literature reviews?)

1. **Narrative Reviews**: It attempts to summarize what has been written on a particular topic but does not seek generalization or cumulative knowledge from what is reviewed.

2. **Descriptive reviews**: They follow a systematic and transparent procedure, including searching, screening and classifying studies. Authors extract from each study certain characteristics of interest, such as publication year, research methods, data collection techniques in the form of frequency analysis to produce quantitative results.

3. **Systematic reviews**: They attempt to aggregate and synthesize in a single source all empirical evidence that meet a set of previously specified eligibility criteria in order to answer a clearly formulated research question.

4. **Realist Reviews** : They are theory-based interpretative reviews designed to enhance traditional systematic reviews by clarifying varied evidence on complex interventions in different contexts, helping to inform policy decisions.

5. **Critical Reviews** : They aim to provide a critical evaluation and interpretive analysis of existing literature on a particular topic of interest to reveal strengths, weaknesses, contradictions, controversies, inconsistencies, and/or other important issues with respect to theories, hypotheses, research methods or results.

2.3 PLAGIARISM AND RESEARCH ETHICS

(Q: What are the conditions of adequacy for understanding plagiarism, and what are its definitions and two key components?)

2.3.1 Conditions of adequacy

It refers to that when creating a definition, it's important to recognize any relevant limitations or restrictions. These restrictions help ensure that the definition is useful and fits well with specific situation. The following criteria for adequacy are relevant to a definition of ''plagiarism'':

•**Fitting language use**: A definition should stay close to how people commonly use language. If definition misses many typical examples of plagiarism, it isn't a good one.

•**Precision**: The greater the precision of the definition, the better it is. For each case, the definition should settle whether or not it is a case of plagiarism.

•**Reliability (intersubjectivity)**: The definition is reliable if different users of it pass the same judgment on specific cases (''If plagiarism is defined as so-and-so, then this is (or is not) a case of plagiarism'').

•**Theoretical fruitfulness**: A good definition is more useful. Ex: a strong definition of plagiarism would not only identify what plagiarism is but also explain why some actions are considered plagiarism while others are not.

•**Relevance for normative purposes**: A good definition of plagiarism should identify actions that people consider morally wrong.

•**Simplicity**: The definition should be homogeneous and ad hoc-free.

2.3.2 Meaning of Plagiarism

Plagiarism refers to misconduct in scientific writing. The basic ideas is that someone deliberately 'copies' someone else's work, in the form of an idea, a method, data, results, or text, and presents it as their own instead of giving credit to the person to whom it belongs to. AS per Merriam-Webster: ''to steal and pass off (the ideas or words of another) as one's own: use (another's production) without crediting the source''.

2.3.3 Two components of plagiarism

Plagiarism is composed of two parts:

(1) **To appropriate the work of someone else**: Plagiarism is stealing someone else's intellectual work such as another's text, tables, graphs, or pictures into one's own paper without having permission to do so. ''To appropriate'' does not have to imply stealing. It could also mean, for instance, acquire, borrow, take, or expropriate.

(2) **Passing it off as one's own by not giving proper credit:** This can be done with or without the approval of the person being plagiarized. If person A uses a passage from a text by B but claims that it was written by C, then, even though it is an incorrect claim, it is not plagiarism, but simply incorrect referring. It is when A claims to have written the passage him- or herself that it becomes plagiarism.

2.3.4 Research Ethics

(Q: Enumerate the key ethical principles that help researchers maintain high ethical standards.)

It provides guidelines for the responsible conduct of research. It educates and monitors researchers to ensure a high ethical standard. Some ethical principles:

1. **Honesty**: Honestly report data, results, methods and procedures, and publication status.
2. **Objectivity**: Avoid bias in experimental design, data analysis, data interpretation, peer review etc.
3.**Integrity**: Keep promises and agreements; act with sincerity; strive for consistency of thought & action.
4. **Carefulness**: Avoid careless errors and negligence; Keep good records of research activities, such as data collection, research design, and correspondence with agencies or journals.
5. **Openness**: Be open to criticism and new ideas.
6. **Respect for Intellectual Property**: Honor patents, copyrights, and other forms of intellectual property by not using unpublished data, methods, or results without permission, giving credit for all contributions to research. Never plagiarize.
7. **Confidentiality**: Protect confidential communications, such as papers/ grants submitted for publication, personnel records, trade or military secrets, and patient records.
8. **Responsible Publication**: Publish in order to advance research and scholarship, not to advance just your own career. Avoid wasteful and duplicative publication.
9. **Responsible Mentoring**: Help to educate, mentor, and advise students. Promote their welfare and allow them to make their own decisions.
10. **Respect for colleagues**: Respect colleagues and treat them fairly.
11. **Social Responsibility**: Promote social good and prevent / mitigate social harms through research, public education, and support.
12. **Non-Discrimination**: Avoid discrimination against colleagues or students on the basis of gender, race, ethnicity, or other factors that are not related to their scientific competence and integrity.
13. **Competence**: Maintain and improve your own professional competence and expertise through lifelong education and learning.
14. **Legality**: Know and obey relevant laws and institutional and governmental policies.
15. **Animal care**: Show proper respect and care for animals when using them in research.
16. **Human Subjects Protection**: When conducting research on human subjects respect human dignity, privacy, and autonomy.
17. "**Other deviations**" from acceptable research practices include:

- Publishing the same paper in two different journals without telling the editors
- Submitting the same paper to different journals without telling the editors
- Not informing a collaborator of your intent to file a patent to make you are the sole inventor
- Including a colleague as an author on a paper in return for a favor
- Discussing with colleagues confidential data from a paper that you are reviewing for a journal
- Trimming outliers from a data set without discussing reasons in paper
- Using an inappropriate statistical technique to enhance the significance of research
- Bypassing the peer review process and announcing results through a press Conference
- Conducting a review of the literature that fails to acknowledge the contributions of other people
- Stretching the truth on a grant application to convince reviewers that your project will make a significant contribution.
- Stretching the truth on a job application or curriculum vita
- Giving the same research project to 2 graduate students in order to see who can do it the fastest
- Overworking, neglecting, or exploiting graduate or post-doctoral students
- Failing to keep good research records
- Failing to maintain research data for a reasonable period of time
- Making derogatory comments and personal attacks in review
- Promising a student a better grade for favors

- Using a racist label in the laboratory
- Making significant deviations from the research protocol approved by Committee
- Not reporting an adverse event in a human research experiment
- Wasting animals in research
- Exposing students and staff to biological risks in violation of biosafety rules
- Rejecting a manuscript for publication without even reading it
- Sabotaging someone's work
- Stealing supplies, books, or data
- Rigging an experiment so you know how it will turn out
- Making unauthorized copies of data, papers, or computer programs
- Deliberately overestimating the clinical significance of a new drug for economic benefits.

2.4 DATA - PREPARING, EXPLORING, EXAMINING AND DISPLAYING

(Q: What is data preparation? what are its benefits? Illustrate the steps that are involved in the process using a block diagram.)

2.4.1 Data preparation:

Data preparation is the activity that ensures the accuracy of the data. It ensures their conversion from raw form to reduced and classified forms that are appropriate for analysis. It includes editing, coding, data entry and preparing a descriptive statistical summary. During this step data entry errors may be revealed and corrected.

2.4.1.1 Benefits of data preparation:

1. To fix errors quickly: Data preparation helps catch errors before processing.
2. To produce top-quality data: Cleaning and reformatting datasets ensures that all data used in analysis will be of high quality.
3. To make better business decisions: Higher-quality data that can be processed and analyzed more quickly and efficiently leads to more timely, efficient, better-quality business decisions.

2.4. Data preparation steps

The fig. 2.4 presents the steps in the data preparation process.

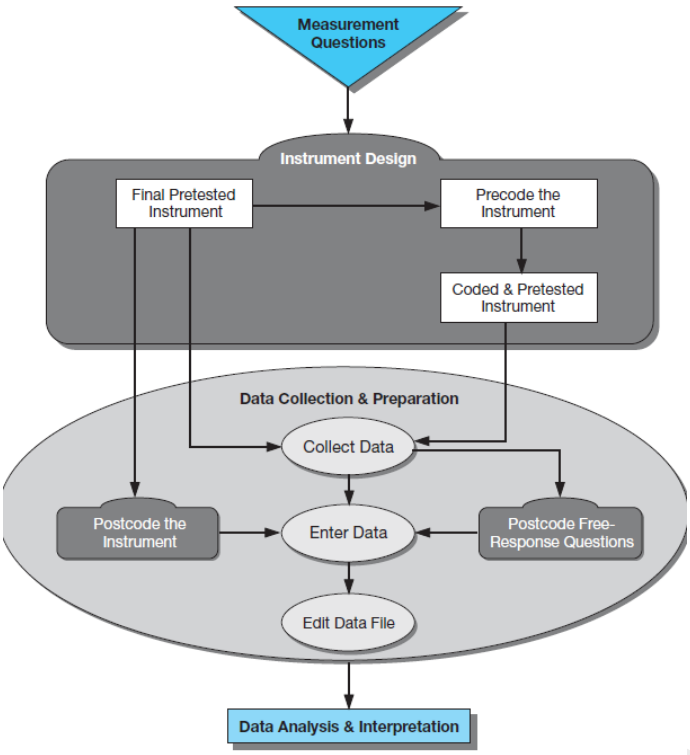


Figure 2.4 Data Preparation in the Research Process

i. Editing

The first step in data preparation is to edit the collected raw data to detect errors and omissions that affect quality. The editor should ensure that the data are accurate, consistent with the intent of the question and other data, uniformly entered, complete and ready for coding.

Survey work includes both:

1. Field Editing: In large projects, the field supervisor is responsible for field editing and should review reporting forms soon after data collection to clarify abbreviations or shorthand, address entry gaps through follow-up calls, to maintain research quality.

2. Central Editing: All data should be carefully edited, with one editor for small studies to ensure consistency, while in larger studies, editing tasks should be divided so each editor focuses on one complete section.

(Q: What should an editor do when encountering obviously incorrect entries in the data, and when is it appropriate to contact the respondent for clarification?
*Q: Explain: **a.** Coding rules. **b.** Spreadsheet data entry. **c.** Bar codes. **d.** Content analysis. e. Missing data. f. Optical mark recognition.*
Q: How should the researcher handle "don't know" responses?)

Editors should correct obvious data errors, like mismatched time units or misplaced entries. Ex: When a respondent clearly specifies time in days (Ex: 13) when it was requested in weeks (Ex: 4 or less) or data is entered in the wrong place or when replies are inappropriate (out of range or not related to the question asked) or missing. They can find the correct answer by reviewing other data, but this should be limited to clear cases. If possible, better to contact the respondent for clarification.

Following **rules guide editors** in their work:

i. Understand the instructions for interviewers and coders.
ii. Do not erase or damage the original responses; they should stay readable.
iii. Use a distinctive color and standard format for all edits in the data.
iv. Initial any changes or added answers.
v. Include initials and the date of editing.

ii. Coding

It is the process of assigning numbers and other symbols to answers so that the responses can be grouped into categories. In coding, categories are the partitions of a data set of a given variable Ex: if the variable is gender, the partitions are male and female. Categorization is the process of using rules to partition a body of data. Both closed- and open-response questions must be coded. Open-ended questions allow

participants to give a free-form text answer. Closed-ended questions restrict participants to one of a limited set of possible answers.

A **codebook**, or **coding scheme**, contains each variable in the study and specifies the application of coding rules to the variable. It is used by the researcher to promote more accurate and more efficient data entry. Codebooks contain the question number, variable name, location of the variable's code on the input medium (e.g., spreadsheet or SPSS data file), descriptors for the response options, and whether the variable is alphabetic or numeric. An example of codebook is shown in fig. 2.5 and Fig. 2.6 shows questions in the sample codebook

Coding Rules: Four rules guide the categorization of a data set:

1. **Appropriateness**: It refers to the suitability of the data for a specific analysis or modeling task.
2. **Exhaustiveness**: Researchers often add an "other" option to a measurement question because they know they cannot anticipate all possible answers. A large number of "other" responses, suggests the measurement scale the researcher designed did not anticipate the full range of information.
3. **Mutual Exclusivity**: Category components should be mutually exclusive. This standard is met when a specific answer can be placed in one and only one cell in a category set. Ex: in a survey, participants are asked for their occupation. Categorization includes (1) professional, (2) managerial, (3) sales, (4) clerical, (5) crafts, (6) operatives, and (7) unemployed. How to code a participant's answer that specified "salesperson at Gap and full-time student"?
4. **Single Dimension:** It refers to a single classificatory principle means every option in the category set is defined in terms of one concept or construct. Ex: the person may be both a 'salesperson' and 'unemployed'. Then the category set should encompass more than one dimension: "occupation type" and "employment status."

Content analysis: Closed questions include scaled items for which answers are anticipated. Ex: rating the satisfaction on a scale from 1 to 5, where 1 is "very dissatisfied" and 5 is "very satisfied."

Question	Variable Number	Code Description	Variable Name
_____	1	Record number	RECNUM
_____	2	Respondent number	RESID
1	3	5 digit zip code 99999 = Missing	ZIP
2	4	2 digit birth year 99 = Missing	BIRTH
3	5	Gender 1 = Male 2 = Female 9 = Missing	GENDER
4	6	Marital status 1 = Married 2 = Widow(er) 3 = Divorced 4 = Separated 5 = Never married 9 = Missing	MARITAL

Fig. 2.5 Sample Codebook of Questionnaire Items

1. What is the zip code of your residence? _ _ _ _ _
2. What is the year of your birth? 19_ _
3. Gender (1) Male (2) Female — Indicate your choice → by number _
4. What is your marital status? (1) Married (2) Widow(er) (3) Divorced (4) Separated (5) Never married — Indicate your choice → by number _
5. Do you own or rent your primary residence? (1) Own (2) Rent (3) Living quarters provided — Indicate your choice → by number _

Fig. 2.6 Sample Questionnaire Items

Open-ended questions are more difficult to code since answers are not prepared in advance. Content analysis is a systematic method for analyzing open-ended questions. It uses preselected sampling units to produce frequency counts and other insights into data patterns. It is described as "a research technique for the objective, systematic, and quantitative description of the manifest content of a communication."

"Don't know" replies: They are evaluated in light of the question's nature and the respondent. While many DKs are legitimate, some result from questions that are ambiguous or from an interviewing situation that is not motivating. It is better to report DKs as a separate category.

Missing data: They occur when respondents skip, refuse to answer, or do not know the answer to a questionnaire item, drop out of the study, or are absent for one or more data collection periods. Researcher error, corrupted data files, and changes to the instrument during administration also produce missing data.
Researchers handle missing data by first exploring the data to discover the nature of the pattern and then selecting a suitable technique for replacing values by deleting cases (or variables) or estimating values.

iii. Data entry

It is accomplished by keyboard entry from precoded instruments, optical scanning, real-time keyboarding, telephone pad data entry, bar codes, voice recognition, OCR, OMR, and data transfers from electronic notebooks and laptop computers.
Database programs, spreadsheets, and editors in statistical software programs offer flexibility for entering, manipulating, and transferring data for analysis, warehousing, and mining.

Data Entry Formats:

a. Keyboarding

Though not efficient, Keyboarding remains a backbone for researchers to create a data file immediately and store it in a minimal space on a variety of media.

i. Database Development

A database is a collection of data organized for computerized retrieval. Database programs allow users to define data fields and link files so that storage, retrieval, and updating are simplified.

ii. Spreadsheet

Spreadsheet is used as a database for data that need organizing, tabulating, and simple statistics. Spreadsheets take numbers, formulas, and text in numbered rows and lettered columns with a matrix of thousands of cells. They also offer some database management, graphics, and presentation capabilities.

b. Optical Recognition

Optical character recognition (OCR) programs transfer printed text into computer files in order to edit and use it without retyping. Optical scanners process the marked-sensed questionnaires and store the answers in a file. Examinees darken small circles, ellipses, or spaces between sets of parallel lines to indicate their answers. Optical mark recognition (OMR) is a more flexible format that uses a spreadsheet-style interface to read and process user-created forms.

c. Voice Recognition

Voice recognition and voice response systems are providing alternatives for the telephone interviewer. Upon getting a voice response to a randomly dialed number, the computer branches into a questionnaire routine.

d. Digital

Using the telephone keypad (touch-tone), an invited participant answers questions by pressing the appropriate number and the computer stores them in a data file.
Bar-code technology: It is used to simplify the interviewer's role as a data recorder. When an interviewer passes a bar-code wand over the appropriate codes, the data are recorded in a small, lightweight unit for translation later.

2.4.2 Exploring, Displaying, and Examining Data

(Q: Differentiate between Exploratory and Confirmatory data analysis.
Q: How can various data visualization methods enhance the understanding of complex datasets? Discuss
Q: With flow diagram explain Data Exploration, Examination, and Analysis)

2.4. Exploratory data analysis (EDA):

It provides a perspective and set of tools to search for clues and patterns in the data. In addition to numerical summaries of location, spread, and shape, EDA emphasizes on visual representations and graphical techniques to provide a complete and accurate impression of distributions and variable relationships.

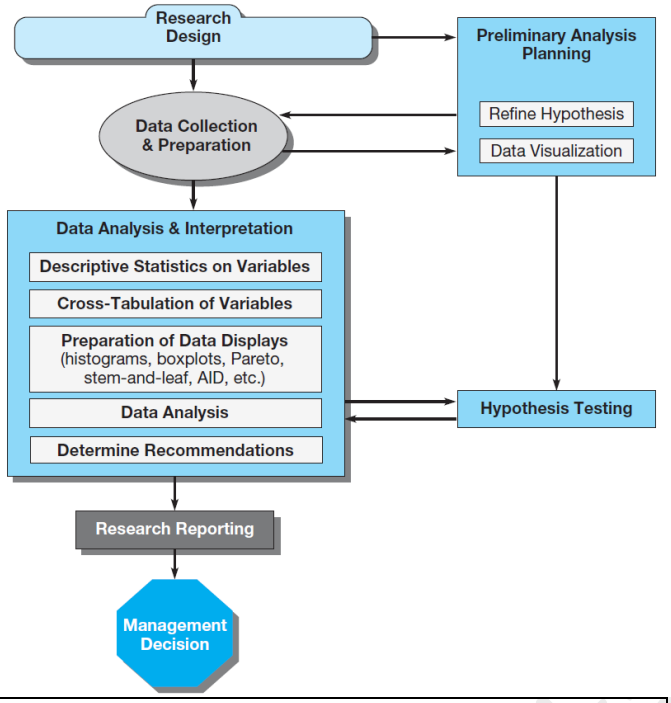


Figure 2.7 Data Exploration, Examination, and Analysis.

Confirmatory data analysis is an analytical process guided by classical statistical inference in its use of significance testing and confidence. Exploratory data analysis is like police detectives searching for clues, as it gathers evidence needed for confirmatory data analysis. Confirmatory analysis, similar to how judges evaluate evidence, uses classical statistics to test significance and confidence. When numerical summaries are used exclusively and accepted without visual inspection, the selection of confirmatory models may be based on flawed assumptions. For these reasons, data analysis should begin with visual inspection. Figure 2.7 presents a flow diagram of Data Exploration. Examination, & Analysis in the Research Process.

Value Label	Value	Frequency	Percent	Valid Percent	Cumulative Percent
21 years old	1	60	6	6	6
18 years old min	2	180	18	18	24
16 years old min	3	330	33	33	57
13 years old min	4	280	28	28	85
10 years old min	5	50	5	5	90
Any age	6	60	6	6	96
No opinion	7	40	4	4	100
		1,000	100	100	

Valid Cases 1,000; Missing Cases 0

Figure 2.8 A Frequency Table (Minimum Age for Social Networking)

Exploratory Data Analysis (EDA) includes following methods for displaying data:

i. Frequency tables

They array data from lowest to highest values with counts and percentages.

They are most useful for inspecting the range of responses and their repeated occurrence.

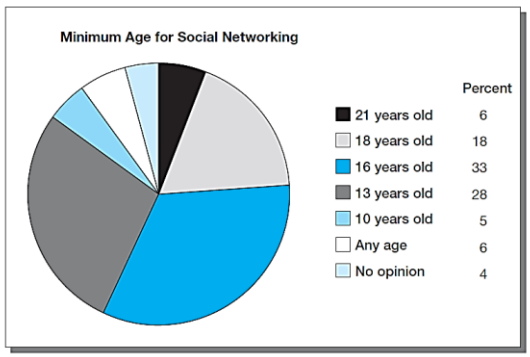
Ex: Figure 2.8 shows a frequency table of the perceived minimum age for owning a social networking account.

ii. Bar charts and pie charts

They are suitable for relative comparisons of nominal data.

Ex: The same data of minimum age for social networking are presented in figure 2.9 using a pie chart and a bar chart. The values and percentages are more readily understood in this graphic format.

i) Pie chart	ii) Bar chart



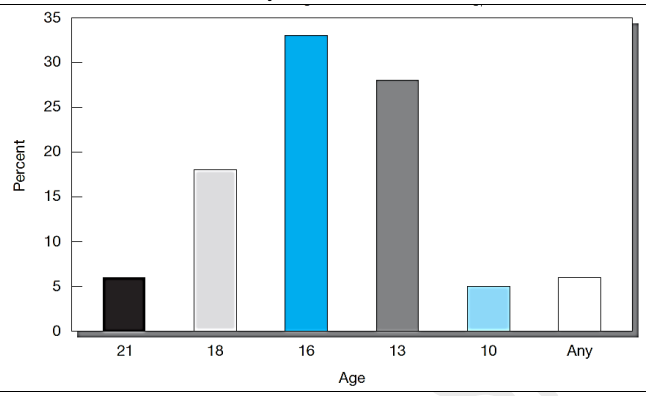


Figure 2.9 Nominal displays of minimum age for social networking data. i) Pie chart ii) Bar chart

iii. Histograms

They are used with continuous variables when it is possible to group the variable's values into intervals. They are useful for (1) displaying all intervals in a distribution, even those without observed values, and (2) examining the shape of the distribution for skewness, kurtosis, and the modal pattern. Histograms can help answer the questions: Is there a single mode? Are subgroups identifiable when multiple modes are present? Are straggling data values (Outliers) detached from the central concentration?

Note: A histogram cannot be used for a nominal variable like minimum age for social networking (Figure 2.5) that has no order to its categories.

Value	Frequency	Percent	Cumulative Percent	Value	Frequency	Percent	Cumulative Percent
54.9	1	2	2	75.6	1	2	54
55.4	1	2	4	76.4	1	2	56
55.6	1	2	6	77.5	1	2	58
56.4	1	2	8	78.9	1	2	60
56.8	1	2	10	80.9	1	2	62
56.9	1	2	12	82.2	1	2	64
57.8	1	2	14	82.5	1	2	66
58.1	1	2	16	86.4	1	2	68
58.2	1	2	18	88.3	1	2	70
58.3	1	2	20	102.5	1	2	72
58.5	1	2	22	104.1	1	2	74
59.9	2	4	26	110.4	1	2	76
61.5	1	2	28	111.9	1	2	78
62.6	1	2	30	118.6	1	2	80
64.8	1	2	32	123.8	1	2	82
66.0	2	4	36	131.2	1	2	84
66.3	1	2	38	140.9	1	2	86
67.6	1	2	40	146.2	1	2	88
69.1	1	2	42	153.2	1	2	90
69.2	1	2	44	163.2	1	2	92
70.5	1	2	46	166.7	1	2	94
72.7	1	2	48	183.2	1	2	96
72.9	1	2	50	206.9	1	2	98
73.5	1	2	52	218.2	1	2	100
				Total	50	100	

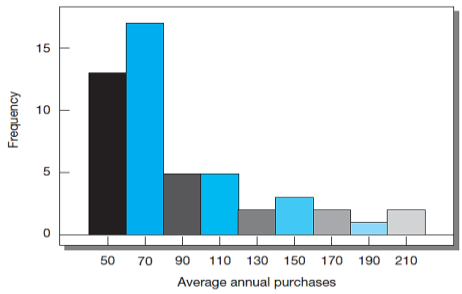


Figure 2.10 Average Annual Purchases of PrimeSell's Top 50 Customers.

Figure 2.11 Histogram of PrimeSell's Top 50 Customers' Average Annual Purchases

Ex: Figure 2.10 gives average annual purchases and figure 2.11 presents a histogram for same. The midpoint for each interval for the variable of interest (average annual purchases) is shown on the horizontal axis; the frequency or number of observations in each interval, on the vertical axis. The height of the bar corresponds with the frequency of observations in the interval.

iv. Stem-and-leaf displays

They are EDA techniques that provide visual representations of distributions. In contrast to histograms, which lose information by grouping data values into intervals, the stem-and-leaf presents actual data values that can be inspected directly.

Adv1: EDA reveals the distribution of values within the interval and preserves their rank order for finding the median, quartiles, and other summary statistics.

Adv2: It eases linking a specific observation back to the data file and to subject that produced it.

5 | 4 5 5 6 6 6 7 8 8 8 8 9 6 | 1 2 4 6 6 7 9 9 7 | 0 2 2 3 5 6 7 8 8 | 0 2 2 6 8 9 | 10 | 2 4 11 | 0 1 8 12 | 3 13 | 1 14 | 0 6 15 | 3 16 | 3 6 17 | 18 | 3 19 | 20 | 6 21 | 8	**Adv3:** Visualization: The range of values and both shape and spread are apparent at a glance. Patterns in the data—such as gaps where no values exist, areas where values are clustered, or outlying values that differ from the main body of the data—are easily observed. Ex: To develop a stem-and-leaf display for the data given in figure 2.10, the first digits of each data item are arranged to the left of a vertical line. Next, pass through the values in the order they were recorded and place the last digit, in unit position, for each item to the right of the vertical line (ignore the decimal values). The stem-and-leaf display is shown in figure 2.12.
Figure 2.12. A Stem-and-Leaf Display of PrimeSell's Average Annual Purchases Data	

v. Pareto diagrams

Pareto diagrams are named after a 19th-century Italian economist and are used in quality management

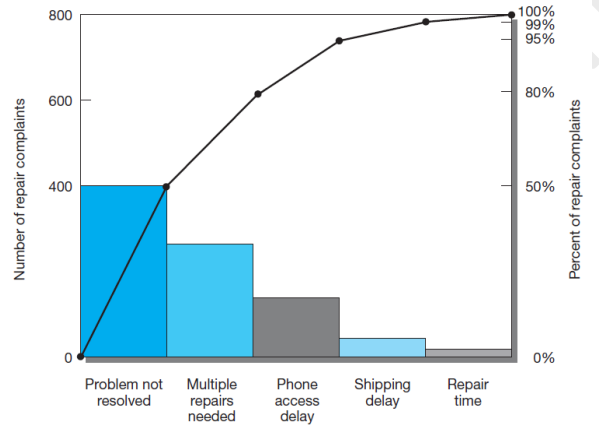


Figure 2.13 Pareto Diagram of MindWriter Repair Complaints

to illustrate the 80/20 rule, where 80% of quality improvements can come from addressing 20% of the causes. In these diagrams, responses are sorted by importance, with bars arranged from tallest to shortest.

Ex: Figure 2.13 shows MindWriter customer complaints, highlighting that the top two issues accounted for 80% of perceptions of inadequate repair service.

vi. Boxplots

Boxplots convey a detailed picture of the distribution's location, spread, shape, tail length, and outliers. They use the five-number summary that consists of the median, the upper and lower quartiles, and the largest and smallest observations.

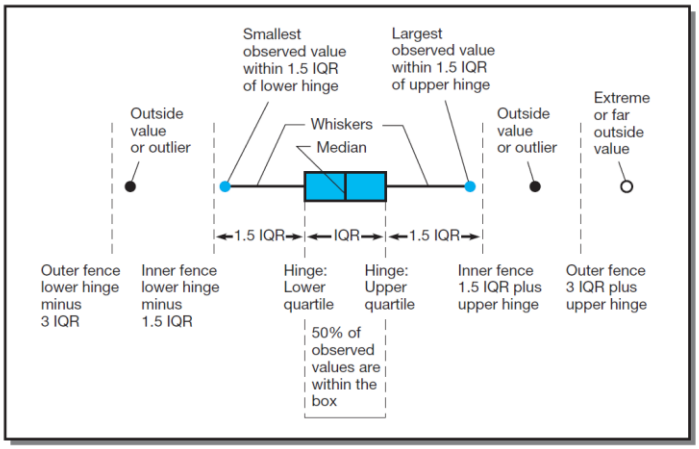


Figure 2.14 Boxplot Components

The basic ingredients of the plot are:
1. The rectangular plot that encompasses 50 % of the data values.
2. A center line marking the median.
3. The edges of the box, called hinges.
4. The "whiskers" that extend from the right and left hinges to the largest and smallest values. These values may be found within 1.5 times the interquartile range (IQR) from either edge of the box. These components and their relationships are shown in Figure 2.14.

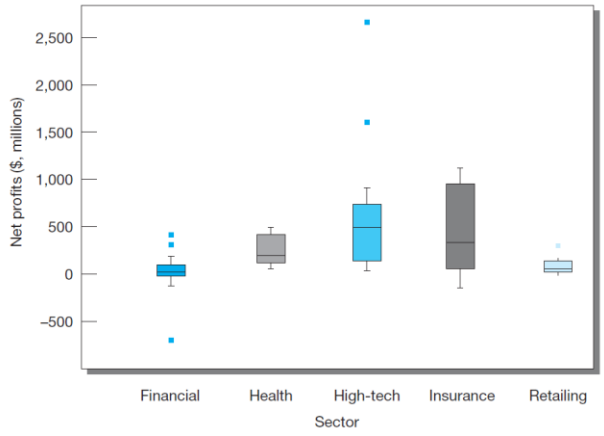


Figure 2.15 Boxplot Comparison of Customer Sectors

Outliers, data points that exceed +1.5 the interquartile range, reflect unusual cases and are an important source of information for the study. Outliers that are entry mistakes should be corrected or removed during editing.

Ex: In figure 2.15, multiple boxplots compare five sectors of PrimeSell's customers by their average annual purchases data.

The overall impression is one of potential problems for the analyst: unequal variances, skewness, and extreme outliers. Note the similarities of the profiles of finance and retailing in contrast to the high-tech and insurance sectors.

vii. Mapping

Geographic Information System (GIS) software and coordinate measuring devices work by linking data sets to each other with at least one common data field (e.g., a household's street address). The GIS allows the researcher to connect target and classification variables from a survey to specific geographic-based databases like U.S. Census data, to develop a richer understanding of the sample's attitudes and behavior. The most common way to display such data is with a map. Colors and patterns denoting knowledge, attitude, behavior, or demographic data arrays are superimposed over street, county, state, or country maps.

2.4.2.2 Cross-tabulation

Cross-tabulation is used to examine relationships involving categorical variables. The tables used for this purpose consist of cells and marginals (row and column total). The cells may contain combinations of count, row, column, and total percentages. It serves as a framework for later statistical testing. Computer software for cross-classification analysis allows for efficient table-based data visualization and decision-

making by incorporating one or more control variables. An advanced variation on *n*-way tables is automatic interaction detection (AID).

References:

1. Ganesan R, Research Methodology for Engineers, MJP Publishers, Chennai. 2011
2. Cooper, Donald R. and Schindler, Pamela S., Business Research Methods, Tata McGraw-hill Publishing Company Limited, New Delhi, India. 2012

(Note: Questions are given for illustrative purpose only)