Irkutsk National Research Technical University
Baikal school of BRICS

09.04.02 Information systems and technology
Program: Information technologies, networks and big data

Course: Data analysis

Atalyan Alina Valerievna
alinaa@mail.ru

What is this course about?

Module 1: Introduction to Data Analysis

Module 2: Data Manipulation and Cleaning

Missing Data (Identifying and dealing with NA values, Imputation techniques)
Outliers (Definition. Outlier Detection. Applications And Techniques)
Data Transformation (Reshaping and aggregating data, Combining datasets)

Module 3: Statistical Analysis

Qualitative and quantitative analysis,
Probability Distributions,
Descriptive Statistics,
Inferential Statistics (Statistical Tests),
Exploratory data analysis (EDA),
Correlation and Regression Analysis (Types of Regression Model, Complex Regression Models)

Module 4: Advanced Data Analysis Techniques

Cluster Analysis
Principal Component Analysis (PCA)
Time Series Analysis

Module 5: Real-World Applications and Case Studies

Analyzing real-world datasets (e.g., healthcare, finance, marketing).
Ethics in Data Analysis
Final Project

Course organization and grading system

>80 points = Represents excellent performance (5)
>70 points = Represents good performance (4)
>60 points = Represents satisfactory performance (3)

Course = Lectures + Practical classes

Theoretical tests (based on lecture materials) or presentations: up to 30 points
Practical classes: up to 50 points
Additional work on data analysis (project): up to 20 points

Penalties:
Absence from classes (starting from 3): minus 2 points each class
Assignment submitted after deadline = 50%

Ideal student

	lecture 1	Practical class 1	lecture 2	Practical class 2	lecture 3	Practical class 3	lecture 4	Practical class 4	lecture 5	Practical class 5	lecture 6	Practical class 6	lecture 7	Practical class 7	test 1 or presentation 1	test 2 or presentation 2	test 3 or presentation 3	task 1	task 2	task 3	task 4	task 5	project	total points	current assessment
Ideal student	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10	10	10	10	10	10	10	10	20	100	excellent

Lecture 1

Introduction to Data Analysis

1. What is data analysis?
    - Definition and key concepts
    - The role of data analysis in business, science, and technology
    - Data life cycle (collection, cleaning, analysis, visualization, interpretation)
2. Data types
    - Structured and unstructured data
    - Qualitative and quantitative data
    - Time series, categorical data, text data
3. Data analysis tools
    - Overview of popular data analysis tools

Definition

The systematic application of statistical and logical techniques to describe the data scope, modularize the data structure, condense the data representation, illustrate via images, tables, and graphs, and evaluate statistical inclinations, probability data, and derive meaningful conclusions known as Data Analysis. [1]

Data analysis is a practice in which raw data is ordered and organized so that useful information can be extracted from it. [2]

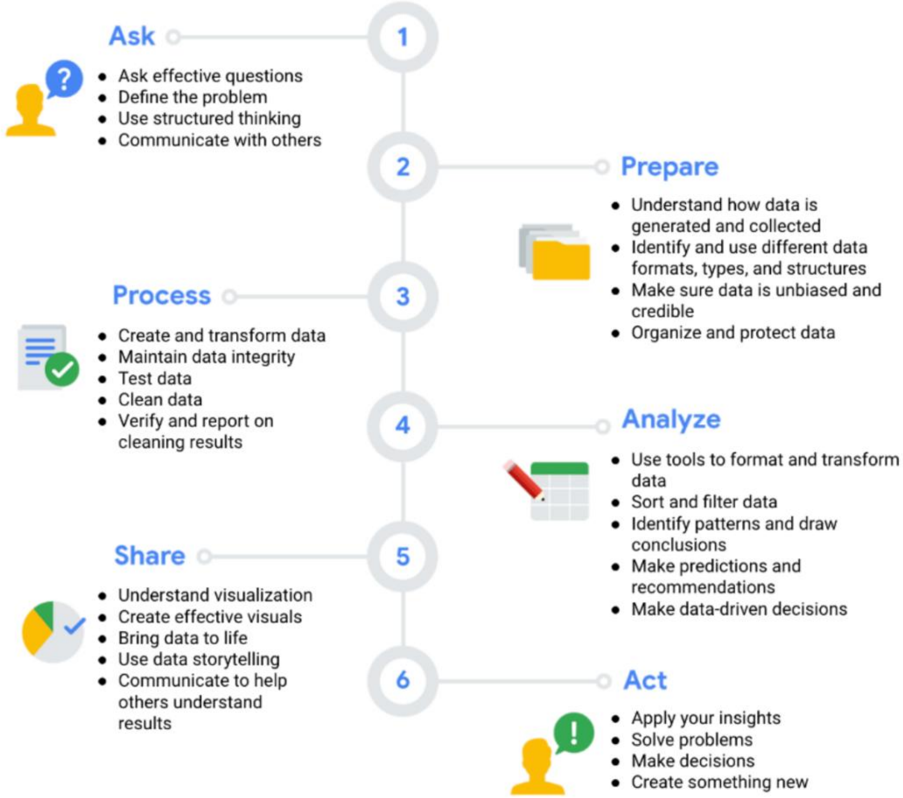
Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making.[3]

[1] Arora, Simran Kaur. "What is data analysis? Methods, techniques & tools." *Retreaved from https://hackr. io/blog/what-is-data-analysis-methods-techniques-tools Last accessed on* 21.03 (2020): 2021.

[2] Tamara Munzner. "Process and Pitfalls in Writing Information Visualization Research Papers". www.cs.ubc.ca. Retrieved 9 April 2018.

[3] Islam, Mohaiminul. "Data analysis: types, process, methods, techniques and tools." *International Journal on Data Science and Technology* 6.1 (2020): 10-15. 10.11648/j.ijdst.20200601.12

Key Concepts[1,2,3] and The Data Analysis Process



Ask 1

- Ask effective questions
- Define the problem
- Use structured thinking
- Communicate with others

2 Prepare

- Understand how data is generated and collected
- Identify and use different data formats, types, and structures
- Make sure data is unbiased and credible
- Organize and protect data

Process 3

- Create and transform data
- Maintain data integrity
- Test data
- Clean data
- Verify and report on cleaning results

4 Analyze

- Use tools to format and transform data
- Sort and filter data
- Identify patterns and draw conclusions
- Make predictions and recommendations
- Make data-driven decisions

Share 5

- Understand visualization
- Create effective visuals
- Bring data to life
- Use data storytelling
- Communicate to help others understand results

6 Act

- Apply your insights
- Solve problems
- Make decisions
- Create something new

[1] Carpineto, Claudio, and Giovanni Romano. *Concept data analysis: Theory and applications*. John Wiley & Sons, 2004.

[2] Reid, Howard M. *Introduction to statistics: Fundamental concepts and procedures of data analysis*. Sage Publications, 2013.

[3] Islam, Mohaiminul. "Data analysis: types, process, methods, techniques and tools." *International Journal on Data Science and Technology* 6.1 (2020): 10-15.

The role of data analysis in business, science, and technology

Data-Driven Decision Making:

Using data to inform and guide business decisions rather than relying solely on intuition or experience.

Examples: Market segmentation, customer behavior analysis, and pricing strategies.

Predictive Analytics:

Using historical data to predict future outcomes.

Examples: Sales forecasting, customer churn prediction, and demand planning

Singh, N., and Amit Kumar Singh. "Data analysis in business research: Key Concepts." *International Journal of Research in Management & Business Studies* 2.1 (2015): 50-55.

Provost, Foster, and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.", 2013.

The role of data analysis in business, science, and technology

Descriptive Analytics:

Summarizing and interpreting historical data to understand what happened.

Examples: Sales performance reports, customer demographics analysis.

Prescriptive Analytics:

Recommending actions based on data analysis to achieve desired outcomes.

Examples: Supply chain optimization, personalized marketing campaigns.

Machine Learning:

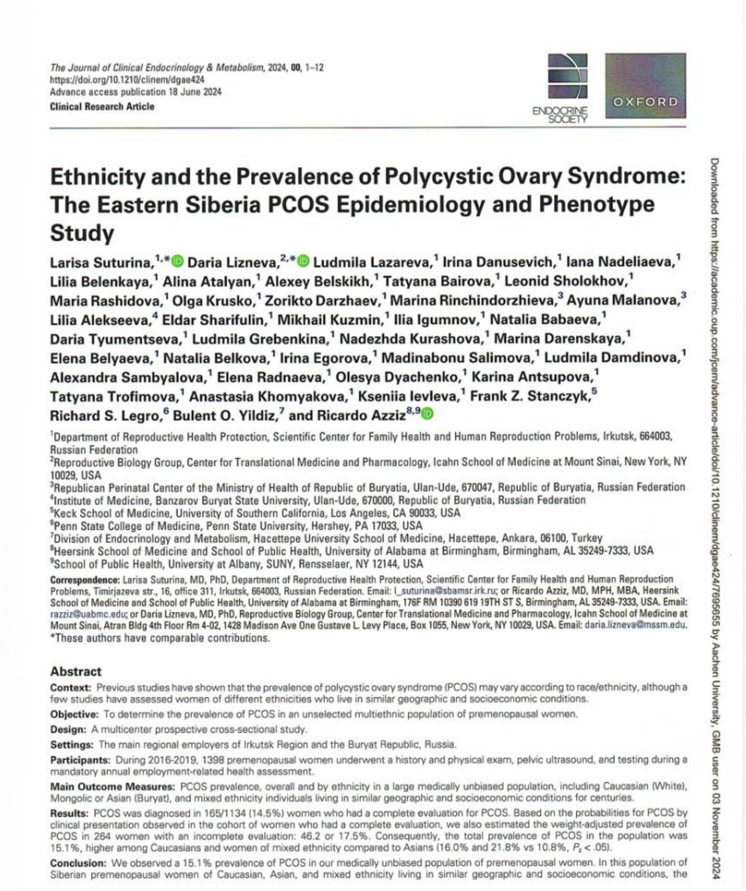
Using algorithms to identify patterns in data and make predictions or decisions without explicit programming.

Examples: Fraud detection, recommendation systems, and sentiment analysis.

Applications for Data Analysis in Research

- Healthcare
  - Example #1: epidemiologists investigate patterns and determinants of disease occurrence and distribution within populations
- Finances
  - Example #2: assessing and managing financial risks
- Environmental studies
  - Example #3: analyze large datasets of temperature records, atmospheric CO2 concentrations, sea level measurements, and other climate variables to detect trends and patterns over time

Example of data analysis of scientific research



The Journal of Clinical Endocrinology & Metabolism, 2024, **00**, 1–12
https://doi.org/10.1210/clinem/dgae424
Advance access publication 18 June 2024
Clinical Research Article

ENDOCRINE SOCIETY

OXFORD

Ethnicity and the Prevalence of Polycystic Ovary Syndrome: The Eastern Siberia PCOS Epidemiology and Phenotype Study

Larisa Suturina,[1,*] Daria Lizneva,[2,*] Ludmila Lazareva,[1] Irina Danusevich,[1] Iana Nadeliaeva,[1] Lilia Belenkaya,[1] Alina Atalyan,[1] Alexey Belskikh,[1] Tatyana Bairova,[1] Leonid Sholokhov,[1] Maria Rashidova,[1] Olga Krusko,[1] Zorikto Darzhaev,[1] Marina Rinchindorzhieva,[3] Ayuna Malanova,[3] Lilia Alekseeva,[4] Eldar Sharifulin,[1] Mikhail Kuzmin,[1] Ilia Igumnov,[1] Natalia Babaeva,[1] Daria Tyumentseva,[1] Ludmila Grebenkina,[1] Nadezhda Kurashova,[1] Marina Darenskaya,[1] Elena Belyaeva,[1] Natalia Belkova,[1] Irina Egorova,[1] Madinabonu Salimova,[1] Ludmila Damdinova,[1] Alexandra Sambyalova,[1] Elena Radnaeva,[1] Olesya Dyachenko,[1] Karina Antsupova,[1] Tatyana Trofimova,[1] Anastasia Khomyakova,[1] Kseniia Ievleva,[1] Frank Z. Stanczyk,[5] Richard S. Legro,[6] Bulent O. Yildiz,[7] and Ricardo Azziz[8,9]

[1]Department of Reproductive Health Protection, Scientific Center for Family Health and Human Reproduction Problems, Irkutsk, 664003, Russian Federation
[2]Reproductive Biology Group, Center for Translational Medicine and Pharmacology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[3]Republican Perinatal Center of the Ministry of Health of Republic of Buryatia, Ulan-Ude, 670047, Republic of Buryatia, Russian Federation
[4]Institute of Medicine, Banzarov Buryat State University, Ulan-Ude, 670000, Republic of Buryatia, Russian Federation
[5]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA
[6]Penn State College of Medicine, Penn State University, Hershey, PA 17033, USA
[7]Division of Endocrinology and Metabolism, Hacettepe University School of Medicine, Hacettepe, Ankara, 06100, Turkey
[8]Heersink School of Medicine and School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35249-7333, USA
[9]School of Public Health, University at Albany, SUNY, Rensselaer, NY 12144, USA

Correspondence: Larisa Suturina, MD, PhD, Department of Reproductive Health Protection, Scientific Center for Family Health and Human Reproduction Problems, Timirjazeva str., 16, office 311, Irkutsk, 664003, Russian Federation. Email: l_suturina@sbamsr.irk.ru; or Ricardo Azziz, MD, MPH, MBA, Heersink School of Medicine and School of Public Health, University of Alabama at Birmingham, 176F RM 10390 619 19TH ST S, Birmingham, AL 35249-7333, USA. Email: razziz@uabmc.edu; or Daria Lizneva, MD, PhD, Reproductive Biology Group, Center for Translational Medicine and Pharmacology, Icahn School of Medicine at Mount Sinai, Atran Bldg 4th Floor Rm 4-02, 1428 Madison Ave One Gustave L. Levy Place, Box 1055, New York, NY 10029, USA. Email: daria.lizneva@mssm.edu.

*These authors have comparable contributions.

Abstract

Context: Previous studies have shown that the prevalence of polycystic ovary syndrome (PCOS) may vary according to race/ethnicity, although a few studies have assessed women of different ethnicities who live in similar geographic and socioeconomic conditions.

Objective: To determine the prevalence of PCOS in an unselected multiethnic population of premenopausal women.

Design: A multicenter prospective cross-sectional study.

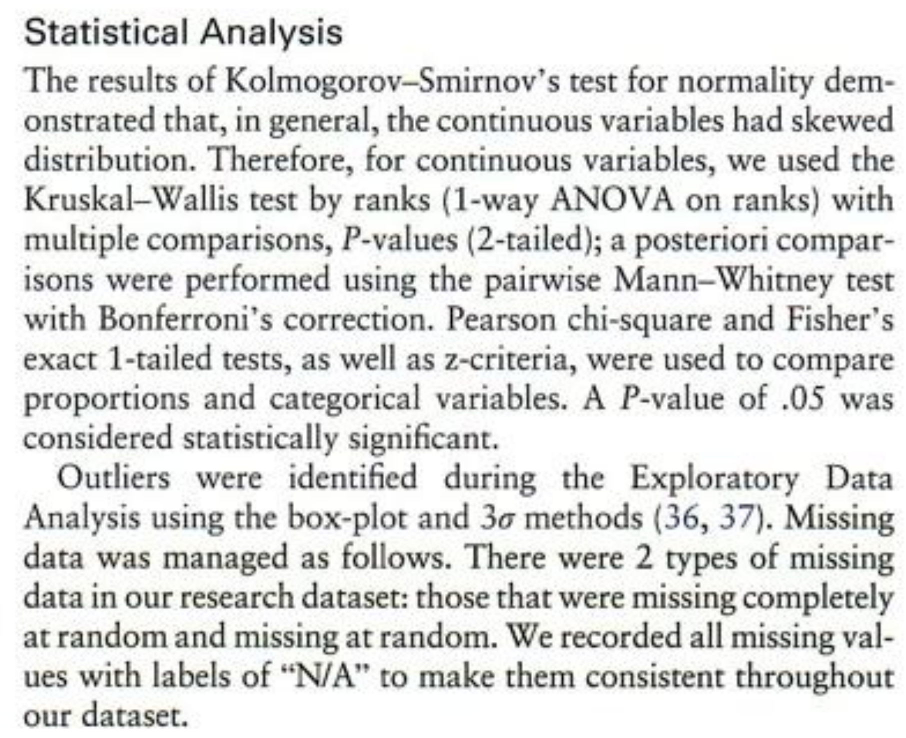
Settings: The main regional employers of Irkutsk Region and the Buryat Republic, Russia.

Participants: During 2016-2019, 1398 premenopausal women underwent a history and physical exam, pelvic ultrasound, and testing during a mandatory annual employment-related health assessment.

Main Outcome Measures: PCOS prevalence, overall and by ethnicity in a large medically unbiased population, including Caucasian (White), Mongolic or Asian (Buryat), and mixed ethnicity individuals living in similar geographic and socioeconomic conditions for centuries.

Results: PCOS was diagnosed in 165/1134 (14.5%) women who had a complete evaluation for PCOS. Based on the probabilities for PCOS by clinical presentation observed in the cohort of women who had a complete evaluation, we also estimated the weight-adjusted prevalence of PCOS in 264 women with an incomplete evaluation: 46.2 or 17.5%. Consequently, the total prevalence of PCOS in the population was 15.1%, higher among Caucasians and women of mixed ethnicity compared to Asians (16.0% and 21.8% vs 10.8%, $P_2 < .05$).

Conclusion: We observed a 15.1% prevalence of PCOS in our medically unbiased population of premenopausal women. In this population of Siberian premenopausal women of Caucasian, Asian, and mixed ethnicity living in similar geographic and socioeconomic conditions, the

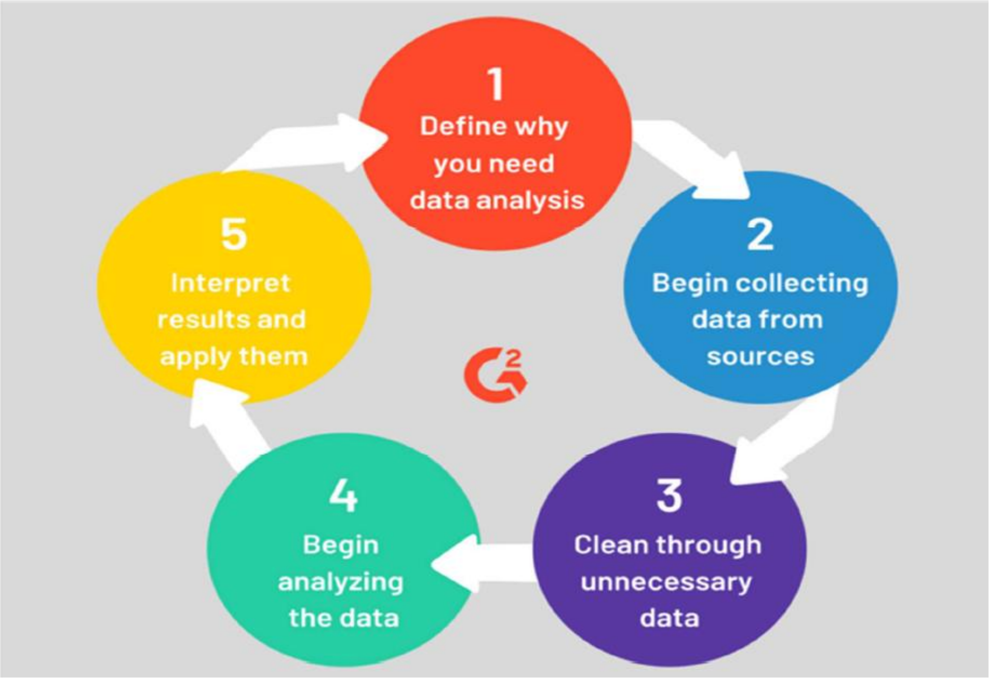


Statistical Analysis

The results of Kolmogorov–Smirnov's test for normality demonstrated that, in general, the continuous variables had skewed distribution. Therefore, for continuous variables, we used the Kruskal–Wallis test by ranks (1-way ANOVA on ranks) with multiple comparisons, *P*-values (2-tailed); a posteriori comparisons were performed using the pairwise Mann–Whitney test with Bonferroni's correction. Pearson chi-square and Fisher's exact 1-tailed tests, as well as z-criteria, were used to compare proportions and categorical variables. A *P*-value of .05 was considered statistically significant.

Outliers were identified during the Exploratory Data Analysis using the box-plot and 3σ methods (36, 37). Missing data was managed as follows. There were 2 types of missing data in our research dataset: those that were missing completely at random and missing at random. We recorded all missing values with labels of "N/A" to make them consistent throughout our dataset.

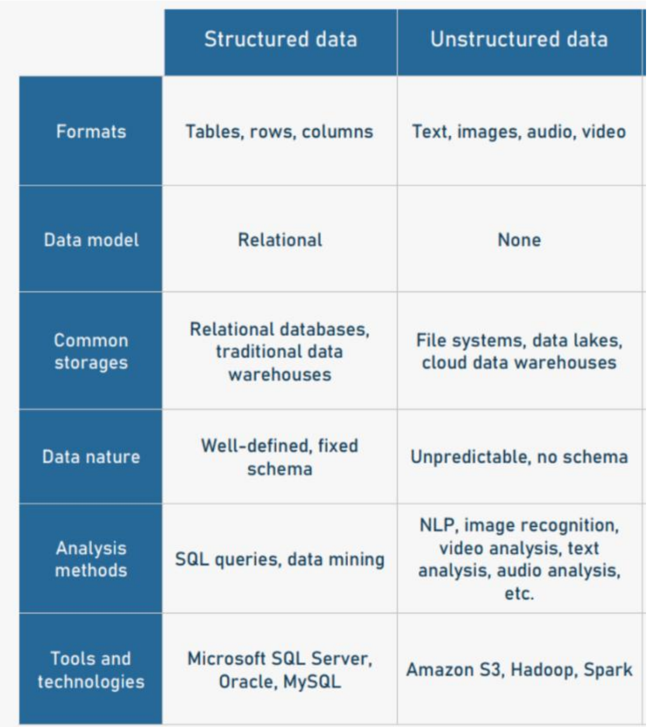
Data life cycle


1
Define why you need data analysis
2
Begin collecting data from sources
3
Clean through unnecessary data
4
Begin analyzing the data
5
Interpret results and apply them

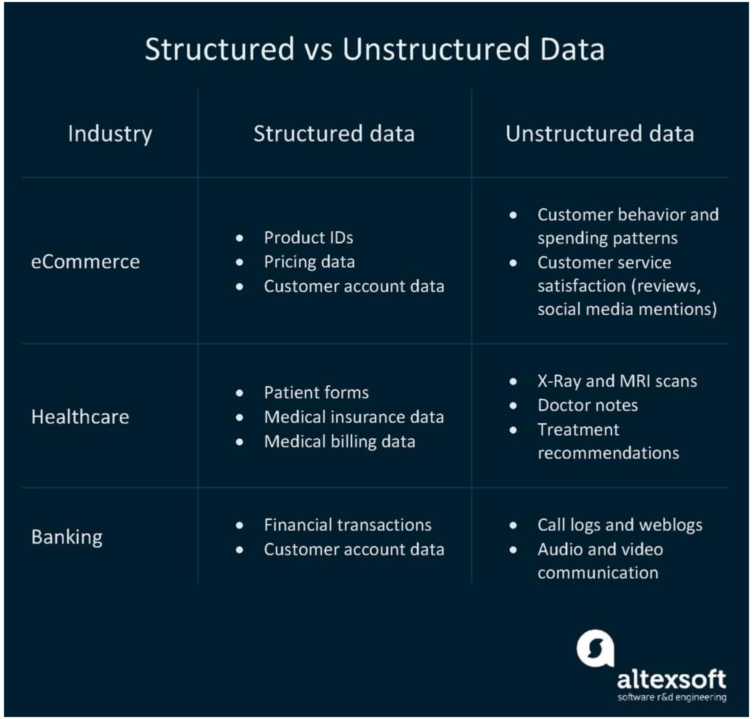
collection,
cleaning,
analysis,
visualization,
interpretation

10.11648/j.ijdst.20200601.12

Structured and unstructured data



	Structured data	Unstructured data
Formats	Tables, rows, columns	Text, images, audio, video
Data model	Relational	None
Common storages	Relational databases, traditional data warehouses	File systems, data lakes, cloud data warehouses
Data nature	Well-defined, fixed schema	Unpredictable, no schema
Analysis methods	SQL queries, data mining	NLP, image recognition, video analysis, text analysis, audio analysis, etc.
Tools and technologies	Microsoft SQL Server, Oracle, MySQL	Amazon S3, Hadoop, Spark

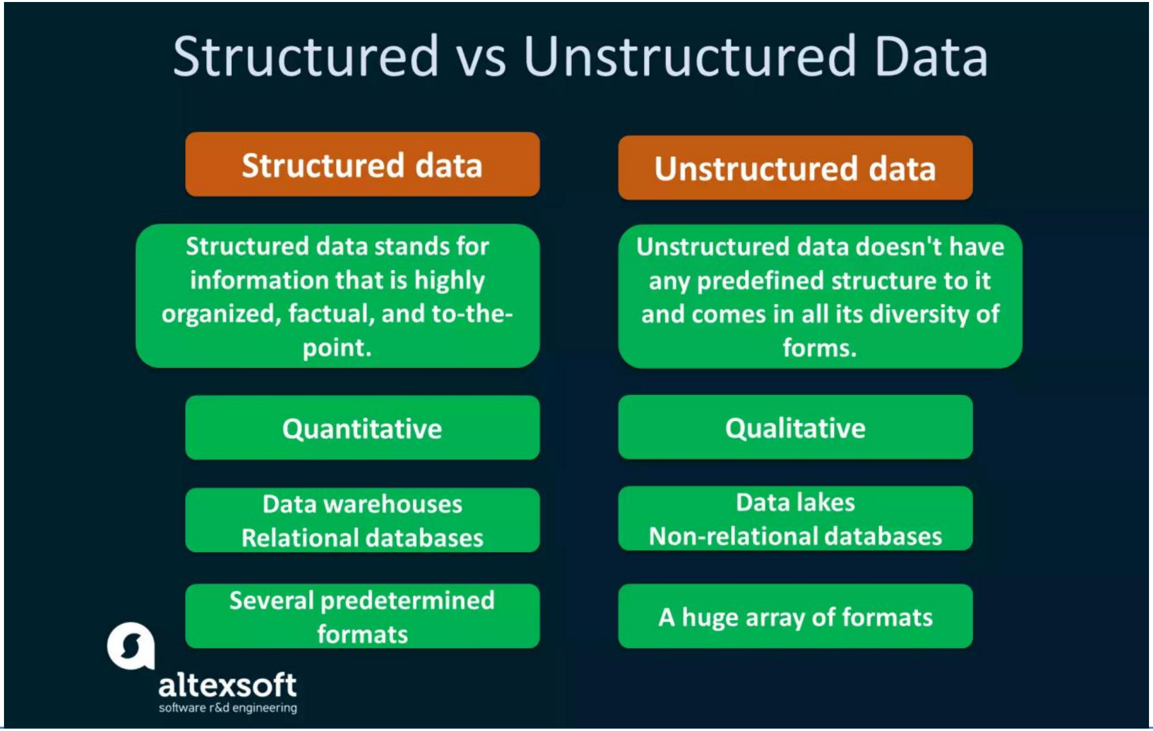


Structured vs Unstructured Data

Industry	Structured data	Unstructured data
eCommerce	• Product IDs • Pricing data • Customer account data	• Customer behavior and spending patterns • Customer service satisfaction (reviews, social media mentions)
Healthcare	• Patient forms • Medical insurance data • Medical billing data	• X-Ray and MRI scans • Doctor notes • Treatment recommendations
Banking	• Financial transactions • Customer account data	• Call logs and weblogs • Audio and video communication

altexsoft
software r&d engineering

Structured and unstructured data

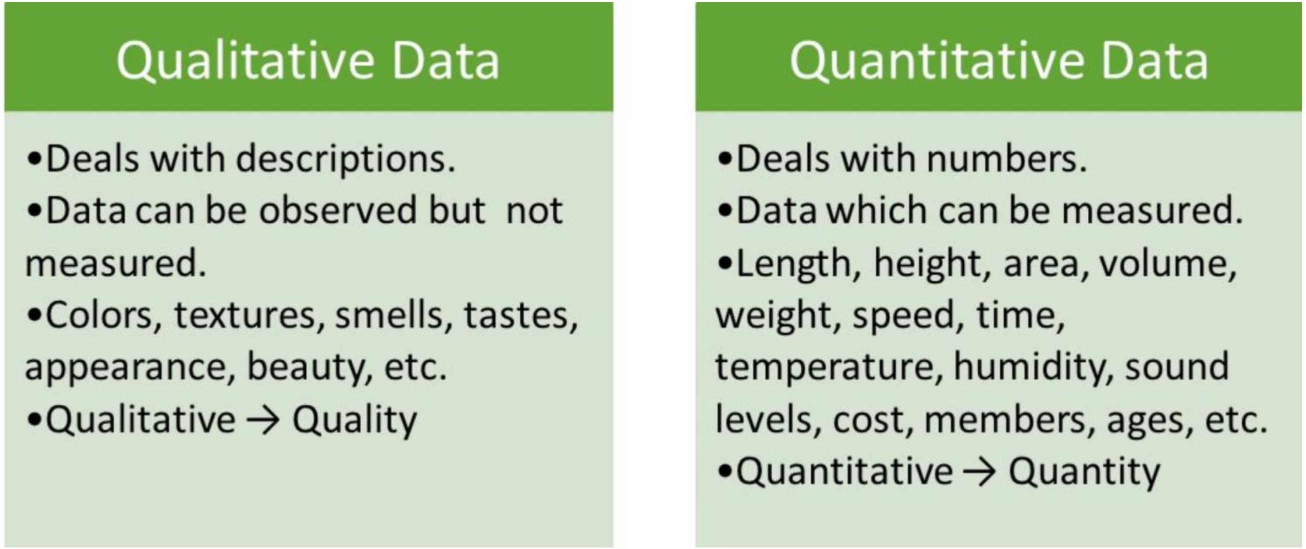


Structured vs Unstructured Data

Structured data	Unstructured data
Structured data stands for information that is highly organized, factual, and to-the-point.	Unstructured data doesn't have any predefined structure to it and comes in all its diversity of forms.
Quantitative	Qualitative
Data warehouses Relational databases	Data lakes Non-relational databases
Several predetermined formats	A huge array of formats

altexsoft
software r&d engineering

Qualitative[1] and quantitative[2,3] data



Qualitative Data

- Deals with descriptions.
- Data can be observed but not measured.
- Colors, textures, smells, tastes, appearance, beauty, etc.
- Qualitative → Quality

Quantitative Data

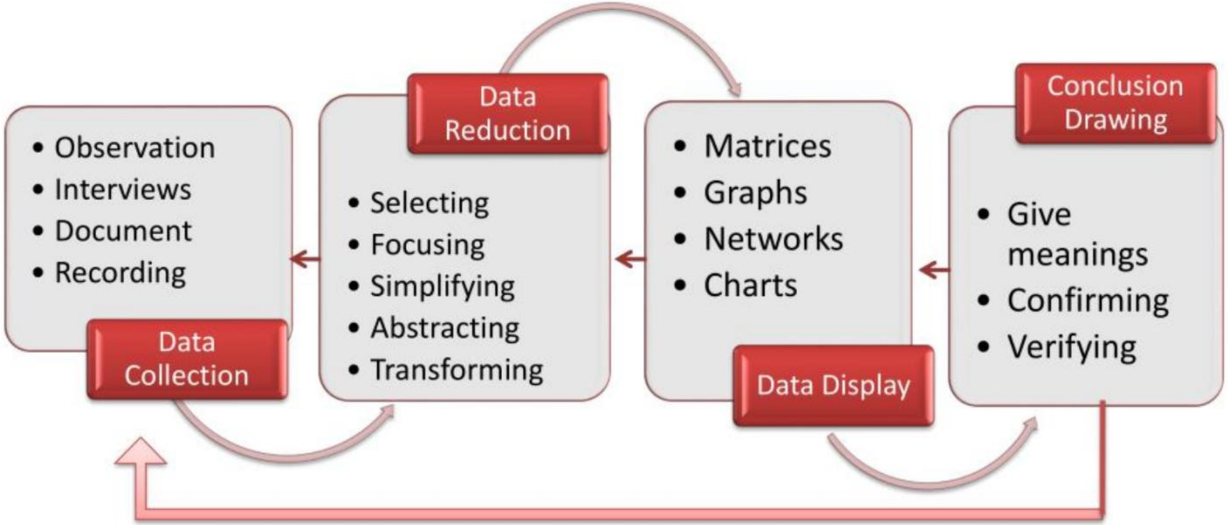
- Deals with numbers.
- Data which can be measured.
- Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc.
- Quantitative → Quantity

[1] Liamputtong, P. (2009), Qualitative data analysis: conceptual and practical considerations. Health Promot J Aust, 20: 133-139. https://doi.org/10.1071/HE09133

[2] Sheard, Judithe. "Quantitative data analysis." *Research Methods: Information, Systems, and Contexts,* (2018): 429-452.

[3] Cramer, Duncan. *Advanced quantitative data analysis*. McGraw-Hill Education (UK), 2003.

Stages in Qualitative Data Analysis



Through analytic processes the researchers turn the data, which is often voluminous, into "a clear, understandable, insightful, trustworthy and even original analysis"

Coding in qualitative data analysis

Coding is the beginning point for most forms of qualitative data analysis.

Coding refers to the process where by researchers define what the data are about, and it is the first step in data analysis.

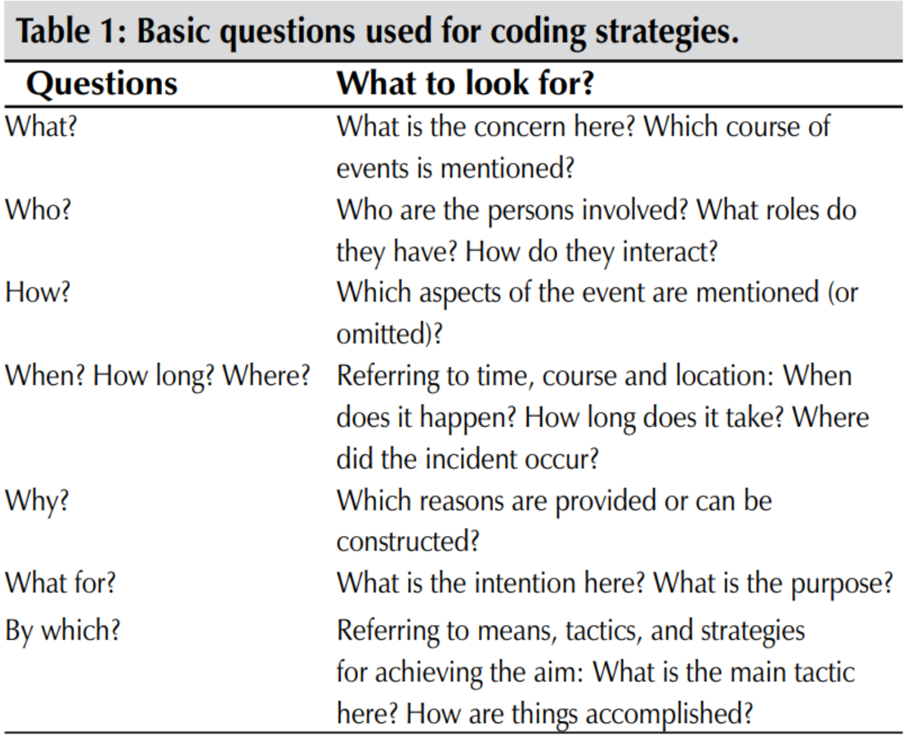


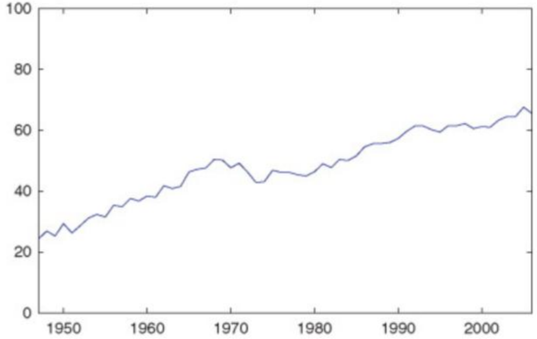
Table 1: Basic questions used for coding strategies.

Questions	What to look for?
What?	What is the concern here? Which course of events is mentioned?
Who?	Who are the persons involved? What roles do they have? How do they interact?
How?	Which aspects of the event are mentioned (or omitted)?
When? How long? Where?	Referring to time, course and location: When does it happen? How long does it take? Where did the incident occur?
Why?	Which reasons are provided or can be constructed?
What for?	What is the intention here? What is the purpose?
By which?	Referring to means, tactics, and strategies for achieving the aim: What is the main tactic here? How are things accomplished?

Liamputtong, Pranee. "Qualitative data analysis: conceptual and practical considerations." *Health promotion journal of Australia* 20.2 (2009): 133-139.

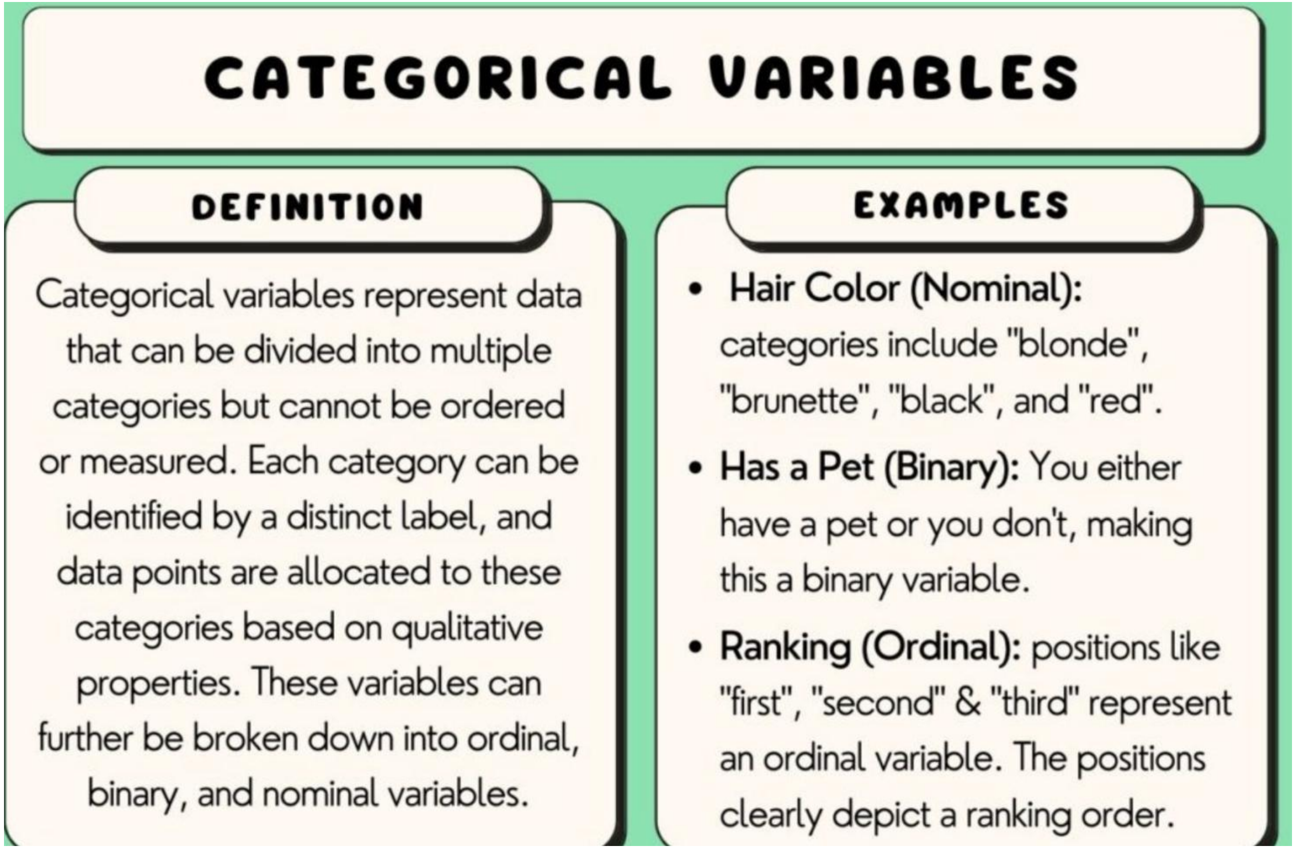
Time series, categorical data, text data

Time series data:
A time series is any sequence of observations recorded at specified times and usually displayed as a time-series plot. This is a graph in which the observations are plotted as a function of time.
Time series abound in all branches of science, engineering, sociology, and economics, and in fact in every field in which observations are recorded over a period of time.



P.J. Brockwell, Time Series Analysis, Editor(s): International Encyclopedia of Education (Third Edition), Elsevier, 2010, https://doi.org/10.1016/B978-0-08-044894-7.01372-5.

Time series, categorical data, text data



CATEGORICAL VARIABLES

DEFINITION

Categorical variables represent data that can be divided into multiple categories but cannot be ordered or measured. Each category can be identified by a distinct label, and data points are allocated to these categories based on qualitative properties. These variables can further be broken down into ordinal, binary, and nominal variables.

EXAMPLES

- **Hair Color (Nominal):** categories include "blonde", "brunette", "black", and "red".
- **Has a Pet (Binary):** You either have a pet or you don't, making this a binary variable.
- **Ranking (Ordinal):** positions like "first", "second" & "third" represent an ordinal variable. The positions clearly depict a ranking order.

Time series, categorical data, text data

Evolution of Data Science Tools

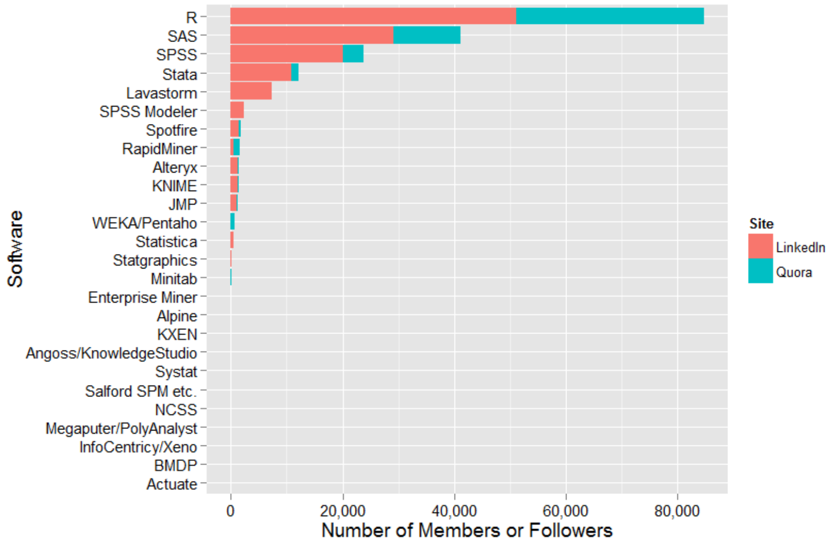
1. Early Statistical and Analytical Tools (1960s-1990s)
2. Emergence of Open Source Programming Languages (2000s)
3. Big Data and Scalable Computing (2010s)
4. Development of Machine Learning Libraries (2010s-Present)
5. Interactive Data Science and Visualization Tools (2010s-Present)
6. AutoML and Cloud-Based Data Science Platforms (2020s)
7. Integrated Data Science Tools (Emerging Trend)

https://www.scaler.com/blog/data-science-tools/

Choosing the Right Tools for Data Analysis

- What is your budget?
- Is there a viable free version, or do you need a subscription?
- What is your technical expertise?
- What is the scalability and flexibility?
- Can the tool integrate with your existing data sources?
- Can it handle the volume of data you’re working with?
- Do you require a tool with modeling capabilities?

Overview of popular data analysis tools



Software	Number of Books
SAS	576
SPSS	339
R	240

The number of books whose titles contain the name of each software package.

«Among the software that tends to be used as a collection of pre-written methods, R, SAS, SPSS, and Stata tend to always be toward the top»

The Popularity of Data Science Software
by Robert A. Muenchen

http://r4stats.com/articles/popularity/

Data analysis tools

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

It is a free language and software for statistical computing and graphics programming.

R is the industry’s leading analytical tool, commonly used in data modeling and statistics.

You can manipulate and present your information readily in various ways.

Topics for presentation on 03/24/2025

1. Missing data (What is missing data? Missing data methods . Evaluation measures).
2. Reshaping and aggregating data (Wide and long format. Summarizing data for reporting, identifying trends, or making data-driven decisions. Examples)
3. Outliers (Definition. Outlier Detection. Applications And Techniques)

~ 10 minutes each presentation

Recommendations for literature search:

1. Determine objective
2. Identify keywords and resource
3. Determine the search strategy

Example:

Objective: *show the evolution of the method in biomedical research*

Resource: PubMed - Free database which includes primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. **pubmed.ncbi.nlm.nih.gov**.

Keywords: *percentile regression, quantile regression, median regression*
Search strategy: *from 1980 (first publication using the method) to 2024*