

What's a Data?

 Information collected for analysis or reference that is often processed and stored on a computer storage medium.

 The basic description of an object, a transaction, or an event, which can be quantitative (numbers) or qualitative (descriptions or categories).

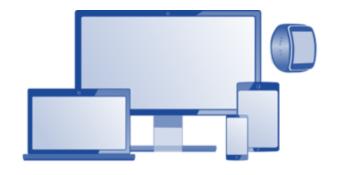
Examples:

| 1 | Car | Red | |
|------|-------|-------|--|
| | | | |
| Sale | Shoes | 167\$ | |



Why is Data Growing?

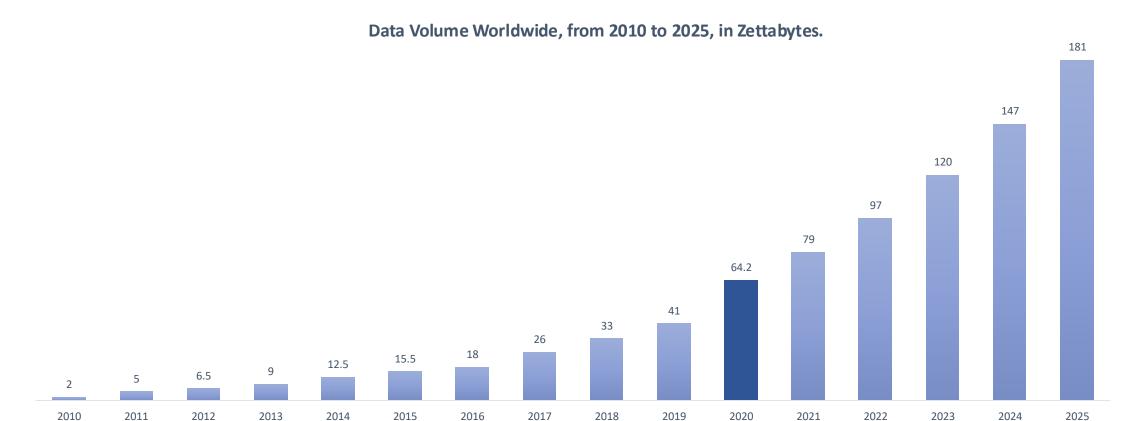
- Technological advances, leading to more and more data being captured, hence Big Data.
- Digitization, Internet of Things (Internet of Bodies), Easier and quicker Internet access, Increased connection speed, Interconnectivity, Portability, Surge in 2020 due to the impact of the pandemic, Etc.







How is Data Growing Across the World?

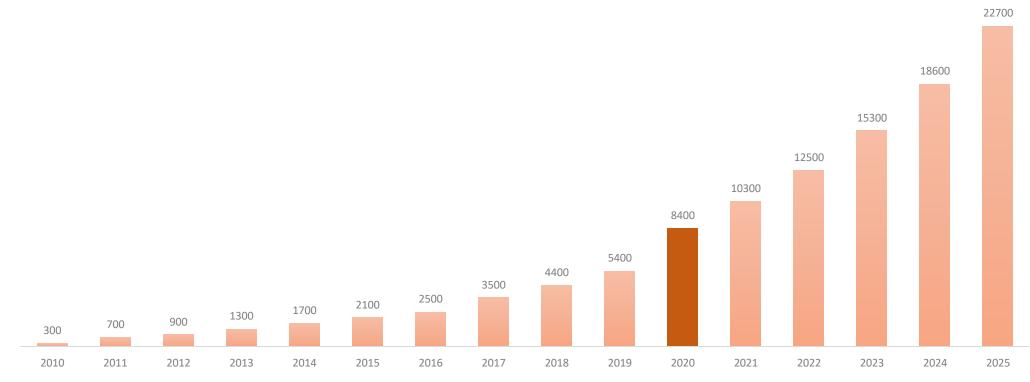


Source of data: Statista

^{*} The data for 2023, 2024, and 2025 are projections.

Worldwide Data Growth Per Capita

Average Estimated Data Volume per Capita, from 2010 to 2025, in Gigabytes.

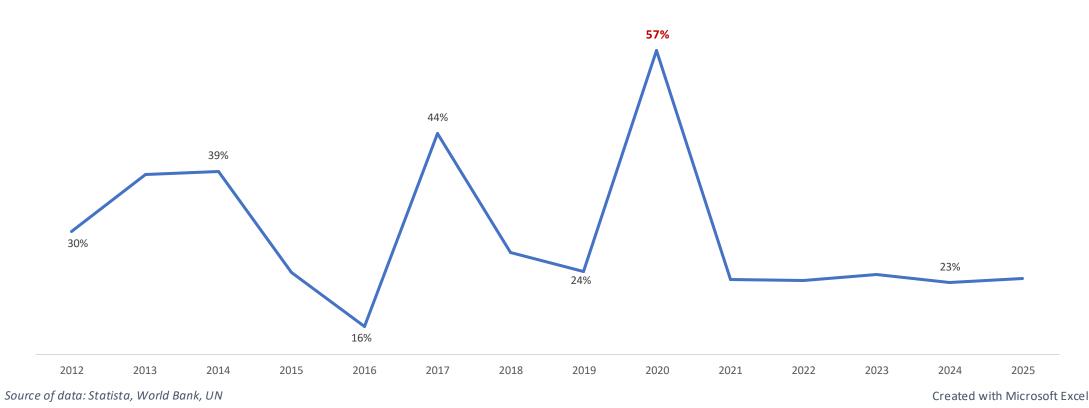


Source of data: Statista, World Bank, UN

Created with Microsoft Excel

More Insights on Data Increase Over Years

Data Volume Increase Worldwide, from 2012 to 2025.



Insights on the Zettabyte

1 zettabyte ≈ 1E⁺¹² gigabytes

Approximately

512 GB USB Universal Serial Bus

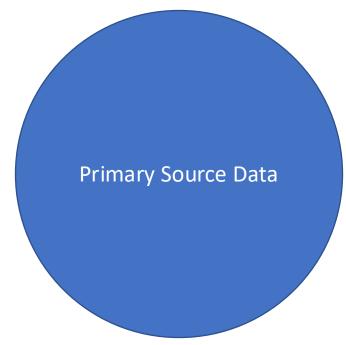
Also close to

16 GB USB Universal Serial Bus

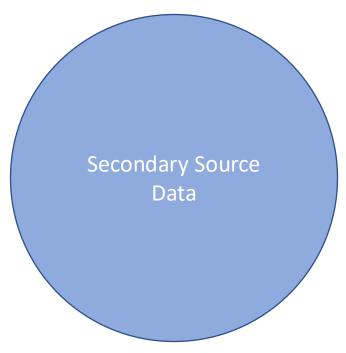
Distribution of 1 zettabyte across the global population in 2025: nearly 8 fully filled 16 GB USB drives per capita.

- Therefore, in 2025, the global total of 181 zettabytes will be equivalent to an average of about 1,400 USB drives of 16 gigabytes per capita, meaning it will be like if each person in the world owned that amount of data.
- These enormous quantities lead to data storage challenges, and research is conducted for DNA data storage. The current largest data quantity measurement unit is the Yottabyte, which is 1024 zettabytes or 1 quadrillion (10¹⁵) gigabytes.

Data Source-Based Classification

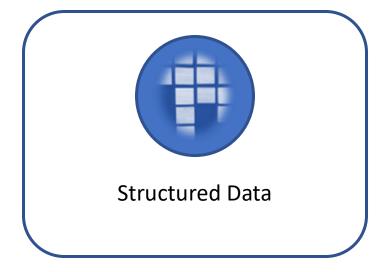


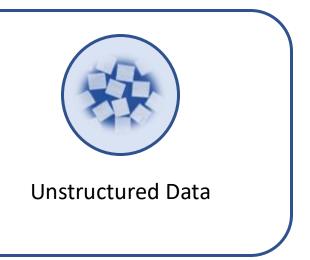
Collected directly via Surveys, Observations, Experience, Questionnaires, Discussions, Etc.

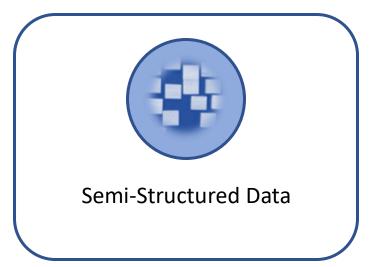


Government Publications, Websites, Books, Journal Articles, Internal Archives, Etc.

Structure-Based Classification







How to Recognize Structured Data?



Layout in a well-defined shape and format

Very well organized according to a very clear structure

Easy to extract and read (via SQL: **Structured** Query Language)

More interdependent and less flexible

Easier to analyze than unstructured or semi-structured data

Wide vs. long formats.

What Does "Wide vs. Long Formats" Mean?



Wide Format

| Deployment | Phones | Tablets | Laptops |
|------------|--------|---------|---------|
| D1 | 43 | 32 | 22 |
| D2 | 29 | 46 | 27 |
| D3 | 37 | 41 | 31 |

- Wide Format: Appropriate for the human brain's comprehension.
- Long Format: Required for some computer programs to be able to perform some tasks. For example, in Python or R, it is more efficient to use this format to create certain types of charts, such as box plots, rather than the wide format.

Long Format

| Deployment | Devices | Quantity |
|------------|---------|----------|
| D1 | Phones | 43 |
| D1 | Tablets | 32 |
| D1 | Laptops | 22 |
| D2 | Phones | 29 |
| D2 | Tablets | 46 |
| D2 | Laptops | 27 |
| D3 | Phones | 37 |
| D3 | Tablets | 41 |
| D3 | Laptops | 31 |
| | | |

Examples of Structured Data



See the tables on the previous slide. And, here is a fun example showing that understanding the information behind the numbers is just as important.

| ID | Employee Name | Job Title | Core Skills | Key Contribution | Favorite Work Perk |
|-----|-----------------|---------------------------------|--|---|--|
| 001 | S. Smiles | Head of Team Happiness | Employee Wellness, Culture Leadership | Launched the "Wellness Wednesdays" program | Weekly mindfulness sessions |
| 002 | P. Productivity | Chief Efficiency Officer | Process Optimization, Time Management | Reduced project turnaround by 30% | Standing desk and flexible work hours |
| 003 | L. Learner | Director of Continuous Learning | Learning & Development, Skill Coaching | Created a mentorship program across teams | Monthly book club on self-development |
| 004 | G. Growth | Innovation Manager | Market Research, Product Development | Led the "Innovation Sprint" for new product lines | Collaboration tools and design sprints |

How to Recognize Unstructured Data?



Completely inorganized Format

Very varied and complex forms

Flexible to analyze

Hard to process due to the lack of a clear structure

Supported by "NoSQL" databases:

- Document Stores MongoDB
- Key-value Stores Redis
- Column-Family Stores Cassandra
- Wide-Column Stores HBase
- Object Stores Amazon S3
- Graph Databases Neo4j.

Examples of NoSQL databases

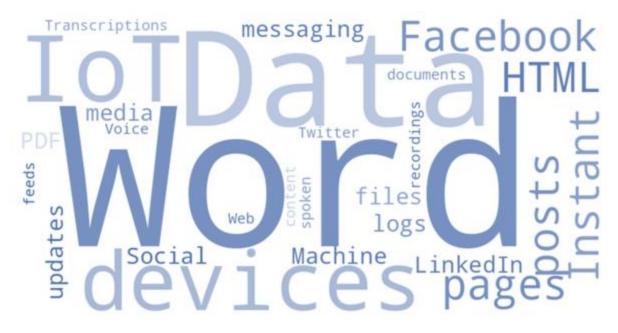
Examples of Unstructured Data



N

THIS

presentation



Word clouds are commonly used for some types of unstructured data (e.g., text data from social media, emails, articles, and reviews) to visually represent the frequency of words in a dataset.

How to Recognize Semi-Structured Data?



Absence of a well-defined formatting, but presence of a certain organization in their structure.

Example of Semi-Structured Data



```
JSON data

(JavaScript Object Notation)

(YAML Ain't Markup Language, originally Yet Another Markup Language)
```

```
"EmployeeID": "12345",

"Name": {

"FirstName": "John",

"LastName": "Doe"

},

"JobTitle": "Software Developer",

"Skills": ["Java", "Python", "SQL"],

"Contact": {

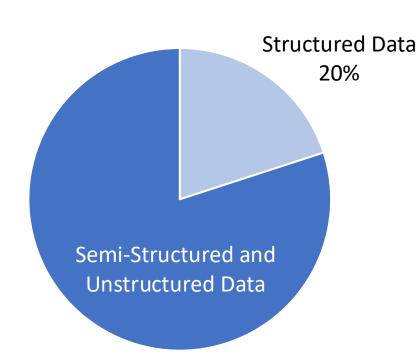
"Email": "johndoe@example.com",

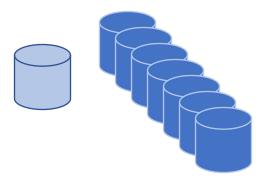
"Phone": "555-1234"

}
```

```
EmployeeID: "12345"
Name:
FirstName: "John"
LastName: "Doe"
JobTitle: "Software Developer"
Skills:
- "Java"
- "Python"
- "SQL"
Contact:
Email: "johndoe@example.com"
Phone: "555-1234"
```

Other Key Aspects of Data





Structured Data require **less storage**, which is not the case for Semi-Structured and Unstructured Data.



Structured Data are easier to process and easier to protect than the Semi-Structured and Unstructured Data.

Usual Data Analysis Applications



New Product Development



Risk Management



New Business Model Development (e.g., Uber, Airbnb)



Strategy (e.g., Growth, Performance.)

Data Types Examples



Consumer Purchase History Claims History of an Insured Driver



Biometric Data



Social Network Data

(Always use them with caution by cross-referencing them with other data)



Meteo Data

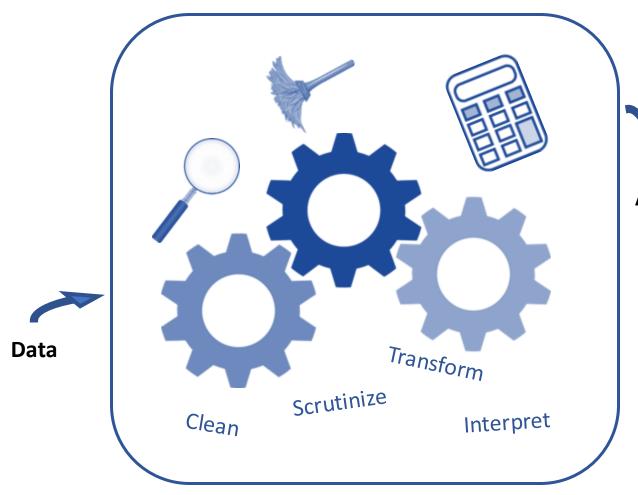


Geolocation Data



Demographic Data.

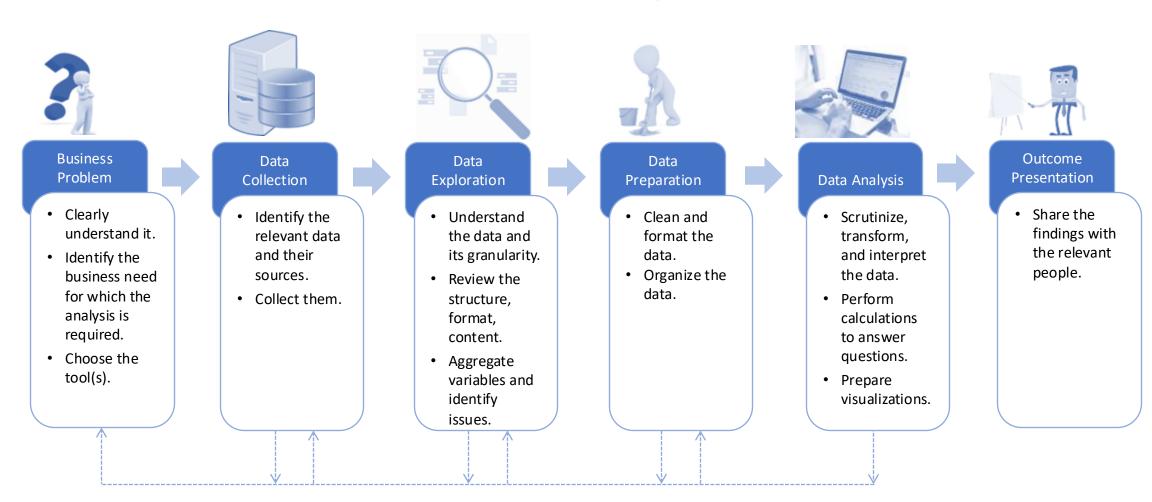
What is Data Analysis?



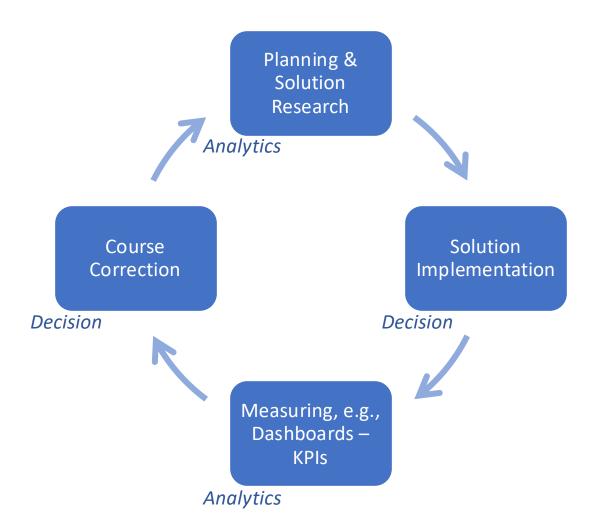
Answers to Questions Informed Decision

The goal is not necessarily to replace traditional decision-making based on experience and gut feeling but to supplement it for greater accuracy and the best possible outcome.

How to Perform Data Analysis?



The Analytic Cycle



- Data analysis is rarely a one-time process; it is usually an iterative approach punctuated with decision-making.
- Once the outcome is shared, an implementation follows, and then an evaluation is conducted. This leads to adjustments in decisions, which often result in another round of data analysis, and so forth.
- This is known as the analytical cycle, in which data analysis is part of the broader analytics process.

How to Clearly Understand the Business Problem?



Know your Audience and their Need.

- Through discussions (across various departments and roles, if needed, e.g., Administration, Finances,
 Operations, Sales, Marketing, IT, etc.).
- This involves communication skills. Even if the first discussion is held with the one who asked for the analysis, other people might have to be considered.
- Choose the contact persons according to the business problem.

Understanding the Business Problem



Pro Tips

- O Visualize the problem from various angles and keep in mind that the view or the approach of a situation can vary widely across roles in the organization.
- Do not neglect the value of the information provided by any speaker, regardless of their role within the business.
- Consider all aspects of the business problem.
- O Do not neglect any question you might think about. If a question comes to your mind, it is better to ask rather than make an assumption.
- After fully understanding the problem, it is ideal to choose the data analysis tool you will use.

How to Choose the Data Analysis Tool?



The tool's choice depends on various factors.

- Type and format of the data you will be working on. Kind of analysis that will be performed and the type of visualizations that will be produced.
- An estimate of the data size and complexity of the analysis. The processing speed. The analyst's skills and experience.
- Integration with other tools (compatibility and collaboration).
- Security (especially for sensitive data, e.g., health). Cost and licensing.

Choosing the Data Analysis Tool

Pro Tips

- For simpler projects and small datasets, tools like Excel or Google Sheets may be sufficient.
- o For more advanced projects, tools like Python, R, or cloud-based platforms like Google BigQuery or AWS are more suitable. Python and R are very versatile tools.
- For big data, Apache Spark, Google BigQuery, Hadoop are ideal.
- Ensure the tool integrates well with your existing data sources (e.g., SQL databases, cloud services, or APIs)
 for smooth import/export of data.
- Decide whether you want built-in visualization capabilities (Tableau, Power BI) or more customization and control (Python - Matplotlib, Seaborn) or R - ggplot2).

How to Identify & Collect the Relevant Data?



Relevant data identification depends on a good understanding of the problem.

- Starting from the problem, determine the type of data that will be needed and its format.
- Target the relevant sources (primary, secondary).
- Sources can be internal or external.
- Knowlegde of data extraction tools and concepts (e.g. SQL vs. NoSQL, ETL vs. ELT, Data Pipelines) is helpful for data accuracy and faster processing.

Identifying & Collecting the Relevant Data



Pro Tips

- Don't overlook any data that might be valuable to your analysis.
- Be flexible. Remain open to adding new data to your collection as your understanding evolves. Although
 the ideal process involves moving through each step once to maintain efficiency, real-world analysis
 sometimes requires backtracking.
- This flexibility works for any other step that comes before presenting the final outcome; you may need to revisit earlier stages. Remaining flexible ensures your analysis stays thorough and relevant, even if it means extra iterations to fine-tune your work.

How to Explore the Data?



Getting familiar with the data.

- Understand the data grain (what a single row or entry in the dataset represents) and validate the data match what we need.
- o Review the structure, format, content.
- Identify discrepencies (missing data, outliers, wrong entries, ...)
- Outline summaries and check central tendancies (sum, count, mean, median,...), patterns and relationships between variables.

Exploring the Data



Pro Tips

- Keep the original data unmodified and work on copies. In Microsoft Excel, format your data as a table from the beginning to unlock powerful features.
- The granularity (level of detail of the data) impacts the insights and size of the data (fine grain -> more detailed insights but increased complexity and larger size; coarse grain -> fewer insights and smaller size).
- o Remain mindful of the audit trails. For example, in Microsoft Excel, use formulas with cell references.
- O Data exploration is like getting to know someone for the first time (the dataset) and asking them to tell you about themself while checking how they fit with your analysis.

How to Prepare the Data?

Refine and optimize the data for accurate analysis.

- Make the data usable, relevant, and reliable for the intended analysis.
- Fix the discrepancies by cleaning the data.
- Handle missing values, inconsistancies, duplicates, incorrect formats and correct these errors.
- Transform and organize the data.
- o Standardize formats, and enrich the data with additional information, if needed.

Preparing the Data



Pro Tips

- Keep in mind your final results and deliverables. For example, anticipate the wording and labeling of the final documents and charts (e.g., tables and chart titles) and name the variables in the dataset accordingly.
- This step is the most time-consuming and can sometimes take up to 80% of the total project time.
- Thorough data preparation smoothes the next steps and minimizes the need to revisit earlier stages, ensuring a more efficient and accurate work.

How to Perform the Analysis?



Uncover insights and test hypotheses.

- Apply statistical methods, algorithms, or models to the refined data to identify patterns, relationships and trends.
- Draw conclusions and find answers to the initial business question.
- Create visual aids (tables, charts, etc.) to showcase your findings clearly and compellingly.

Performing the Analysis



Pro Tips

- Think of this step as assembling a puzzle. Now that all the pieces are in place, from the preparation phase,
 you now connect them to reveal the big picture.
- You now derive what the data tells you about the business question at hand.

Some Types of Analyzes

Descriptive

What happened.

Summarizes and describes the main features of a dataset.

Exploratory

What patterns might exist.

Explains the underlying causes of past events or outcomes.

Inferential

What the data suggests.

Takes a small sample of a group and stretches its implications to the larger population.

Predictive

What could happen.

Forecasts future outcomes based on historical data.

Diagnostic

Why it happened.

Investigates the data to discover patterns, relationships, or anomalies.

Prescriptive

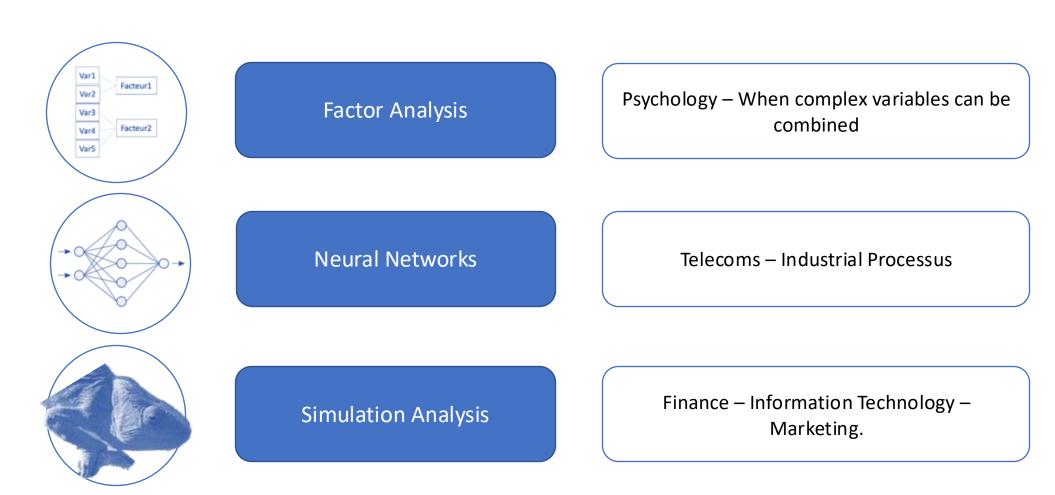
What should be done.

Combines the previous analyses to suggest actionable strategies and decisions.

Some Analysis Technics (1)



Some Analysis Technics (2)



Using Microsoft Excel to Perform the Analysis

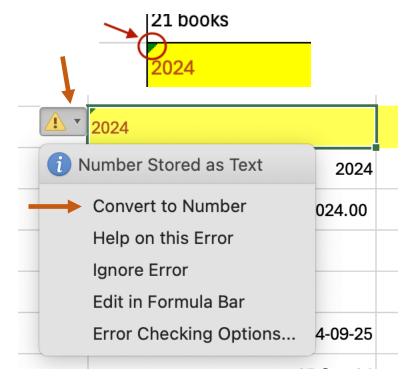
Excel can be used in the following situations:

- small to medium-sized datasets
- Simple and straightforward analysis needs
- Basic Data Exploration and Cleaning (Sorting and Filtering, Removing duplicates, Handling missing data)
- Basic summaries (mean, median, mode, standard deviation, etc.)
- Pivot Tables and Conditional formatting
- Interactive dashboards
- Data Analysis Toolpak, What-If Analysis (Scenario Manager, Goal Seek, Data Table).
- Avoid using Excel for large datasets (hundreds of thousands to millions of rows) due to performance limitations. Excel also struggles with complex analysis and is unsuitable for extensive statistical tasks. For such purposes, a more robust tool like R, Python, or SAS is preferable.

Preventing Errors Caused by Numbers Stored as Text

Calculation errors can derive from numbers stored in cells as text. Checking the cells' contents' default alignment can be essential to identify and fix these cases.

| В | С | D |
|------------------------------|--------------------------------------|--|
| Examples | Type (Based on Default Alignment) | Calculations Outcome |
| Chapter 7 | Text | #VALUE! |
| Data Analysis | Text | #VALUE! |
| 21 books | Text | #VALUE! |
| 2024 | <u>Text</u> | Throws error, except in recent versions. |
| 2024 | Number | Gives a result |
| 2,024.00 | Number | |
| 2024-09-28 | Date | |
| 28-Sep-24 | Date | |
| Saturday, September 28, 2024 | Date | |
| TRUE | Boolean | |
| FALSE | Boolean | |



This can also be fixed using the "VALUE" or "NUMBERVALUE functions."

Useful Functions to Explore, Prepare & Analyze the Data (1)

ROUND

Rounds a number to a specified number of digits.

RIGHT

Returns the last character or characters in a text string, from the position you specify.

EXACT

Tells if two text strings are exactly the same.

VALUE

Converts a text string that represents a number to a number.

LEFT

Returns the first character or characters in a text string, up to the position you specify.

CLEAN

Removes all non-printable characters.

TEXT

Lets you change the way a number appears by applying formatting to it with format codes.

MID

Returns a specific number of characters from a text string, starting at the position you specify.

TRIM

Removes all spaces from text except for single spaces between words.

Useful Functions to Explore, Prepare & Analyze the Data (2)

LEN

Returns the number of characters in a text string.

LOWER

Converts letters in a text string to lowercase.

SUBSTITUTE

Substitutes a text for another in a text string to replace specific text in the text string.

FILTER

Filters a range of data based on criteria you define.

UPPER

Converts all letters in a text to uppercase.

REPLACE

Replaces part of a string with a different string based on the number of characters you specify.

SORT

Sorts the contents of a range.

PROPER

Capitalizes the first letter and any letters that follow any character other than a letter.

CONCATENATE

Joins two or more text strings into one string.

Useful Functions to Explore, Prepare & Analyze the Data (3)

FIND

Locate one text string within a second text string.

LOOKUP

Look in a single row or column and find a value from the same position in a second row or column.

XLOOKUP

Searches a range or an array, and then returns the item corresponding to the first match it finds.

INDEX

Returns a value or the reference to a value from within a range.

HLOOKUP

Searches for a value in the top row of a table or array, and then returns a value in the same column from a row you specify in the table or array.

COUNT(A)

COUNT counts the number of cells that contain numbers (COUNTA counts the number of cells that are not empty).

MATCH

Returns the relative position of a specified value within a range.

VLOOKUP

Searches for a value in a column of a table or array and returns a value in the same row from a column you specify in the table or array.

COUNTIF(S)

COUNTIF counts the number of cells that meet a criterion (COUNTIFS applies many criteria across multiple ranges and counts the number of times all criteria are met).

Useful Functions to Explore, Prepare & Analyze the Data (4)

IF

Compares values with what you expect.

SUM

Adds values.

SUMIF(S)

Sums values in a range that meet criteria that you specify (multiple criteria for SUMIFS).

AVERAGE

Returns the average (arithmetic mean) of the arguments.

AVERAGEIF(S)

Returns the average (arithmetic mean) of all the cells in a range that meet a given criteria (multiple criteria for AVERAGEIFS).

MEDIAN

Returns the median of the given numbers.

MODE

Returns the most frequently occurring, or repetitive, value in an array or range of data.

MIN

Returns the smallest number in a set of values.

MAX

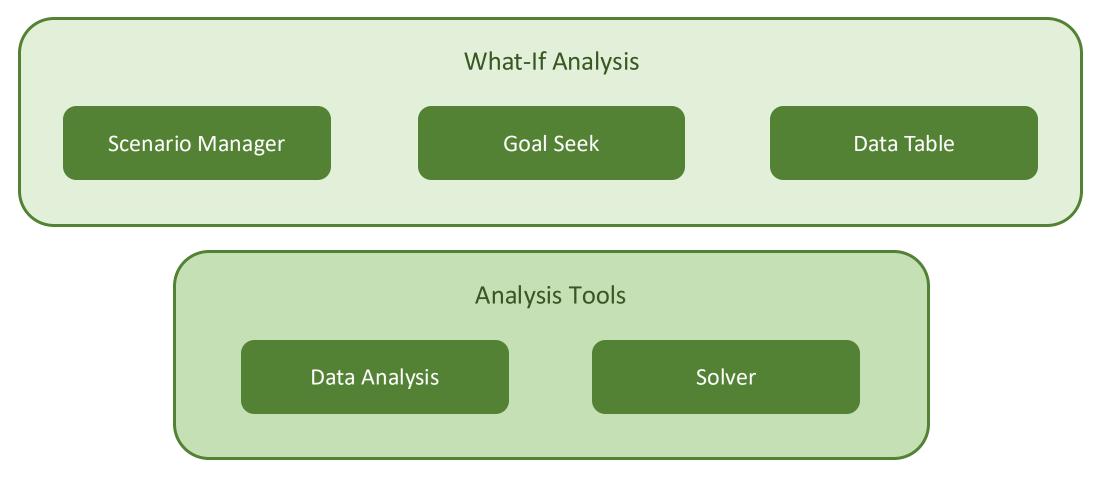
Returns the largest number in a set of values.

Useful & Advanced Features

Remove Duplicates Text to Columns Name Manager **Data Validation** Consolidate Pivot Table **Conditional Formating Power Query** Power Pivot **Power View**

For more information and demos about a feature, click on the corresponding box.

Other Useful Tools in the Data Tab



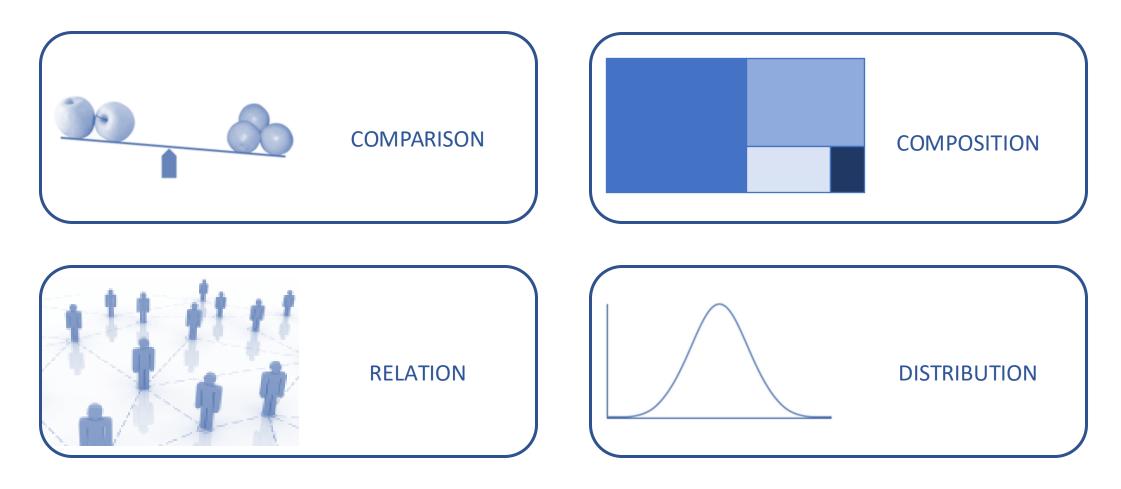
For more information and demos about a tool, click on the corresponding box.

Creating Visualizations

Communicate a message in simple, straightforward and universal language: an image.

- o It is highly risky to perform data analysis without visualizations. They make it easier to spot patterns, trends, outliers, and relationships in the data. Without them, important insights may be overlooked.
- O Visualizations help in identifying errors or inconsistencies in the data for further investigation. Without them, errors may remain undetected, leading to faulty analysis and decision-making.
- They also help to communicate data findings clearly to stakeholders. Complex data presented without visualizations can lead to misinterpretation, bias or incomplete interpretations, poor communication, and ultimately misguided business decisions.

Goal of the Visualization



How to Choose the Right Visualization?

Choosing the right visualization depends on the type of data you want to present and the message you want to convey. Below are some key questions that can help with this task.

How many variables do you want to display?

This helps to determine whether you want to use single-variable charts (e.g., bar or line charts) or multi-variable charts (e.g., scatter plots, bubble charts, or stacked bar charts). The number of variables influences the chart's complexity.

Do the variables have a hierarchy?

This auestion if important you consider charts like treemaps or hierarchical bar charts. If there's a hierarchy, charts that reflect lavers of categories or subcategories could be a good choice.

Are the variables related or comparable in nature?

This helps determine whether to use comparison charts such as grouped bar charts or scatter plots.

Are the variables ordered?

This is highly relevant to deciding whether to use line or bar charts reflecting a sequence or ranking (e.g., time series).

How to Increase Visualization's Quality?

Pro Tips

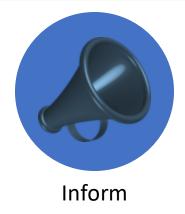
- Know the data well and understand its type (categorical, numerical, time series, etc.).
- Avoid overloading and keep the visualization clear and straightforward. In Microsoft Excel, lines and grids
 are often added by default. Consider removing them and anything you don't need for enhanced clarity
 and quality. Minimize the use of numerical data in the charts. If labels are displayed, the y-axis and other
 extraneous information can be removed to avoid redundancy and streamline your visuals.
- O Use colors for specific purposes, not randomly. Avoid default colors. For example, start with gray as a base and add colors according to intentions. Do not use more than 5 colors. Be mindful of color blindness (For example, work with nuances of Orange/Blue instead of Red/Green).
- Examples of charts are found in pages 3, 4 and 5.

Sharing the Findings through a Presentation



Delivering the answers to the business question.

- Good preparation and communication skills are essential.
- O Defining the purpose of the presentation is key. Your goal impacts how you deliver your message and can be one of, or a combination of, the following:







Inspire & Motivate

Planning & Preparing the Presentation

Know your audience and your purpose.

Identify the audience and what they hope for from your presentation to tailor your content to their interests and knowledge level. This allows you to make your presentation more engaging and effective and guides its direction.

Develop the structure of the presentation.

Create a clear outline that organizes the content logically, like a narrative or story, including an introduction, main points (the findings), and a conclusion.

Plan how you end the presentation.

Conclude with a strong summary of your key points and a call to action or closing thought. A memorable ending reinforces your message and encourages your audience to reflect on what they've learned.

Plan how you will start.

Your opening should grab attention and set the tone for the presentation. Consider starting with a relevant anecdote, a surprising fact, or a thought-provoking question to engage your audience right from the beginning.

Prepare visual aids.

Visual aids (slides, charts, graphs) enhance understanding and retention of information. Ensure your visuals are clear, relevant, and professional. They should complement your verbal message rather than overwhelm it.

Anticipate questions and rehearse.

Consider possible questions your audience may have and prepare your responses. Rehearsing your presentation will build your confidence, help you manage your time, and refine your delivery.

Delivering the Insights

- The presentation marks a critical moment. It encapsulates all the work done since the beginning, mostly completed for this important step.
- As the analyst, you have been privileged to access insights from the data. It's as if you've had a conversation with it, allowing it to speak to you in a language that not everyone can understand.
 Now, it is time to translate the message and share this precious information provided by the data with stakeholders, empowering them to make informed decisions that drive success.
- In addition to the quality of the analysis and the visual aids, how you convey this information is also critical, and aspects such as posture, voice tone, and audience engagement should be carefully considered.

Tips for an Effective Delivery (1)

Set the scene 2 **Engage Immediately** 3 Be Clear and Concise Use Visual Aids Effectively 5 Stay Structured 6 **Emphasize Key Points** Be Prepared for Questions

Begin with a brief introduction to set the context. State your name, role (*if needed*), and the purpose of the presentation to establish rapport with the audience.

Start with a compelling statistic, a thought-provoking question, or a relevant story related to your data to capture your audience's attention from the outset. Encourage interaction during the presentation.

Use clear language and avoid technical jargon unless you are sure all stakeholders understand it. If necessary, explain any complex terms briefly.

Focus on visuals (charts and graphs) to convey your message clearly and avoid cluttered slides. Limit the amount of text on each slide and avoid reading it.

Follow a logical flow (introduction, analysis, findings, and conclusion) to help your audience follow your thought process.

Use verbal emphasis, such as changing your tone or pausing, to draw attention to key insights and findings. Reiterate important points to ensure they stick with the audience.

Questions foster engagement and show that you value stakeholders' input. Be open to them and, at the beginning, clearly let your audience know whether you will welcome their questions during the presentation or at the end.

Tips for an Effective Delivery (2)

Practice Active 8 Listening Use Body Language 9 Effectively Remain Flexible 10 11 Conclude Strongly **Provide Supporting** 12 Materials End with a Call to 13 Action 14 Reflect and Improve

When stakeholders ask questions or provide feedback, listen carefully. Acknowledge their points and respond thoughtfully to demonstrate your understanding.

Use confident body language – stand tall, make eye contact, and use gestures to emphasize points. This helps convey confidence and keeps the audience engaged.

Be prepared to adjust your presentation based on audience reactions or questions. If something resonates particularly well or sparks interest, explore that topic further.

Summarize your main findings and their implications in your conclusion. Offer actionable recommendations that stakeholders can take away.

Offer handouts or supplementary materials after (or before) your presentation for stakeholders to review later. This reinforces your message and provides additional value.

Clearly outline the next steps or actions you expect from your stakeholders following the presentation. This encourages follow-up and collaboration.

After the presentation, ask for feedback from trusted colleagues or stakeholders about what worked well and what could be improved. Use this to enhance your future presentations.

More About Questions and Interactions

Questions during the presentation, rather than at the end, can be highly helpful to foster Asking engagement. questions or prompting discussions throughout the presentation promotes a collaborative atmosphere. However, this can also be disruptive. If you choose to answer questions during the presentation, ensure you don't lose track of the structure you set up.

When answering questions, if you don't have an answer on hand, take note and let the person who asked it know that you will get the answer. Then, after the presentation, find them the answer and provide it within three days, if possible. Avoid leaving unanswered questions, and also avoid promising an answer without providing it.

After the presentation, be open to feedback and questions. Offer to share your slides and any additional resources, reinforcing your commitment to the audience's understanding.

- o In a world increasingly dominated by technology and with vast amounts of data being continuously captured, there is an opportunity for more informed, effective, and accurate decision-making. If a business is likened to a vehicle, and the decision-maker is the driver, then data analysis serves as light, empowering the driver for a safe and smooth journey.
- As technology advances, numerous tools with useful features and enhanced capabilities emerge.
 While many data analysis tools have surpassed Microsoft Excel in various functionalities, it remains a strong, versatile, and accessible contender in the field.
- Furthermore, isn't it interesting that Microsoft Excel's Power Query, Power Pivot, and Power View features can be considered the ancestors of Power BI, a robust data analysis tool?

I hope you enjoyed this data journey.





