Olivier Maugain

Data Literacy

Course Notes

365 DataScience

TABLE OF CONTENTS

ABSTRACT

Being data literate means having the necessary competencies to work with data. Any manager or business executive worth their salt is able to articulate a problem that can be solved using data.

If you want to build a successful career in any industry, acquiring full data literacy should certainly be one of your key objectives.

Someone who is data-literate would have the ability to:

- Articulate a problem that can potentially be solved using data
- Understand the data sources involved
- Check the adequacy and fitness of data involved
- Interpret the results of an analysis and extract insights
- Make decisions based on the insights
- Explain the value generated with a use case

The course is organized into four main chapters.

First, you will start with understanding data terminology – we will discuss the different types of data, data storage systems, and the technical tools needed to analyze data.

Then, we will proceed with showing you how to use data. We'll talk about Business Intelligence (BI), Artificial Intelligence (AI), as well as various machine and deep learning techniques.

In the third chapter of the course, you will learn how to comprehend data, perform data quality assessments, and read major statistics (measures of central tendency and measures of spread).

We conclude this course with an extensive section dedicated to interpreting data. You will become familiar with fundamental analysis techniques such as correlation, simple linear regression (what r-squared and p-values indicate), forecasting, statistical tests, and many more.

Keywords: data, data literacy, data storage, data analysis, statistics

1. Introduction

1.1. What Exactly Is Data Literacy

Definition: "Data literacy is the ability to read, understand, create, and communicate data as information."

A data literate person has the necessary competencies to work with data:

- Articulate a problem that can potentially be solved using data
- Understand the data sources used
- Check the adequacy and fitness of the data involved
- Interpret the results of an analysis and extract insights
- Make decisions based on the insights
- Explain the value generated with a use case

Some of the most important questions a data literate person should be able to answer are:

- How do we store data?
- Which are the systems we use to do that?
- Are the data complete and clean enough to support a correct decision?
- What are the main characteristics of a data set?
- What methodology was applied to analyze the data?
- How reliable is the result of an analysis or forecast?

1.2. Why Do We Need Data Literacy

The concept of data literacy continues to grow in popularity. Reasons:

- Customers leave their digital footprints when engaging with products and brands, such as website clicks, app registration, mobile devices, e-commerce purchases, social media behavior, and even physical store visits
- Patients' data are collected by physicians during the treatment of their conditions.
- Video data continue to be collected by CCTV and physical stores
- Even cars are on the verge of becoming autonomous, thanks to the huge amount of data analyzed by automotive firms and the advancements in machine and deep learning
- Devices and machines are producing and processing more and more data. Most people feel overwhelmed with the flood of data and information, and do not know how to deal with it

Data literacy helps organizations make sense of all their data, creating value and:

- Better customer understanding
- Faster decision making
- More accurate predictions
- Optimizing activities
- Reducing risks and costs

- Improving productivity
- Better serving customers, suppliers, patients, etc.

2018 Survey "Lead with Data - How to Drive Data Literacy in the Enterprise" by the software vendor Qlik

Results:

- 24% of business decision makers are fully confident in their own data literacy
- 32% of senior leaders are viewed as data literate
- Only 21% of 16 to 24-year-olds are data literate
- 94% of respondents using data in their current role recognize that data help them do their jobs better
- 82% of respondents believe that greater data literacy gives them stronger professional credibility

Conclusion:

While enterprise-wide data literacy is considered important, data literacy levels remain low.

2020 Survey "The Human Impact of Data Literacy" by Accenture and Qlik

Results:

- Data-driven organizations benefit from improved corporate performance, leading to an increased enterprise value of 3-5 percent
- Only 32% of companies are able to realize tangible and measurable value from data

- Only 21% of the global workforce are fully confident in their data literacy skills
- At the same time, 74% of employees feel overwhelmed or unhappy when working with data

Conclusion:

Although employees are expected to become self-sufficient with data and make data-driven decisions, many do not have sufficient skills to work with data comfortably and confidently.

1.3. Data-Driven Decision Making

Gut-feeling decisions:

- It can contribute to the development of solutions
- People have a poor history of previous decisions
- The human mind forgets, ignores, or rejects times you were wrong, while it remembers when gut feeling turned out to be correct
- Gut feeling can be blind to key input if the relevant facts are not concrete or objective enough
- "Gut feeling" seeks a "good feeling", ignoring important factual information

Definition:

"Confirmation bias is the tendency to search for, interpret, favor, and recall information in a way that confirms or supports one's prior beliefs or values."

It leads people to unconsciously select the information that supports their views while dismissing non-supportive information. Confirmation bias is not the only cognitive bias the human mind is exposed to.

Data-driven decisions:

- They constitute evidence of past results; with data, you have a record of what works best, which can be revisited, examined, or scrapped if useless
- Data and their analysis allow us to get more information
- A large pool of evidence is the starting point for a virtuous cycle that builds over time and may benefit not only a single decision-maker, but the organization as a whole
- The analysis of data allows us to handle more attributes, values, parameters, and conditions than the human mind could process

Conclusion:

Gut feeling should not be the only basis for decision making. Data should not be overlooked, but rather added to the mix of ideas, intuition, and experience when making decisions

1.4. Benefits of Data Literacy

Benefits of working with data:

- Get useful insights
- Guide in solving a problem
- Inform one's own decisions

More and more professionals will be expected to be able to use data in their roles.

An entrepreneur or business owner can use data to determine:

- Who are the most loyal clients?
- What are the key cost drivers for the business?
- Which posts had the strongest impact in the last social media campaign?
- What kind of users visit the company's website and are more likely to convert?

An **employee** or middle manager in an enterprise can use data to:

- Tweak a marketing campaign and increase its effectiveness
- Calculate which media platforms or ad formats yield the highest ROIs
- Convince executives to make a certain investment
- Or analyze revenue performance

Important notice:

The use of data is not about weakening humans and their decision-making power. It should not be seen as a threat to managers' jobs. Instead, data support and to inform them to make better decisions.

Conclusion:

To be successful in a "digital world", one needs to become more confident in the use of data.

1.5. How to Get Started

There are four stages of data literacy. This course concerns the first stage only.

First stage:

- Terminology: Understand and employ the right terms
- Describe and communicate with data
- Ascertain the quality of data
- Interpret and question the results of analyses performed by others
- Extract information: Understand and internalize the importance of data for one's own success

Second stage:

- Prepare data
- Choose the right chart to visualize information
- Carry out complete analyses
- Identify patterns and extract insights
- Tell business stories using data: Become autonomous in the processing of data and the extraction of value from these

Third stage:

- Design analyses and experiments
- Understand statistical concepts, and apply the right techniques and tools
- Remember and observe the assumptions and conditions related to the techniques employed: Put data into operation into one's own business domain (e.g., marketing, medicine, sports, social science, etc.)

Fourth stage:

- Develop and fine-tune statistical or mathematical models
- Choose the right machine learning algorithm
- Apply scientific standards in the use of data
- Interpret the output of a predictive model to make sure that the results are reliable
- Using programming language: Data as a profession

2. Understanding Data

2.1. Data Definition

Definition:

"Data are defined as factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation."

Data should be used in plural; the singular form is datum, which is a single value of a single variable.

Data ≠ Information

- Data are the raw facts
- Information is derived from data
- Data need to be processed and interpreted in a certain context in order to be transformed into information

Examples of data:

- A Spreadsheet with sales information
- E-mails
- Browsing history
- Video files you shared on social media
- Geolocation as tracked by mobile phones
- The amount of fuel consumption recorded by vehicles

Types of data:

- Quantitative vs. qualitative data
- Structured vs. unstructured data
- Data at rest vs. data in motion
- Transactional vs. master data
- (Small) data vs. “Big” data

2.2. Types of Data

2.2.1. Qualitative vs. Quantitative Data

Quantitative data:

Data that can be measured in numerical form. The value is measured in the form of numbers or counts. Each data set is associated with a unique numerical value. Quantitative data are used to describe numeric variables. Types of quantitative data:

- Discrete data: Data that can only take certain values (counts). It involves integers (positive and negative)

  Examples:

  - Sales volume (in units)
  - Website traffic (number of visitors, sessions)
- Continuous data: Data that can take any value. It involves real numbers.

  Examples:

  - ROI (return of investment) of a project
  - Stock price

- Interval (scale) data: Data that are measured along a scale. Each point on that scale is placed at an equal distance (interval) from one another, with no absolute zero. Examples:
  - Credit score (300-850)
  - Year of the foundation of a company
- Ratio (scale) data: Data that are measured along a scale with an equal ratio between each measurement and an absolute zero (the point of origin). They cannot be negative.

  Examples:
  - Revenue
  - Age

Qualitative data:

Data that is collected in a non-numerical form. It is descriptive and involves text or categories, but also integers (when recoded). Types of qualitative data:

- *Nominal (scale) data*: Data that do not have a natural order or ranking. They cannot be measured. Calculations with these data are meaningless.

  Examples:
  - Marital status
  - Response to an email campaign (yes/no)
- *Ordinal (scale) data*: Data that have ordered categories; the distances between one another are not known. Order matters but not the difference between values.

Examples:

- Socio-economic class (“poor”, “working class”, “lower middle class”, “upper middle class”, “upper class”)
- Product rating on an online store (number of stars, from 1 to 5)

- *Dichotomous data*: Qualitative data with binary categories. They can only take two values, typically 0 or 1. Nominal or ordinal data with more than two categories can be converted into dichotomous data. Instead of X possible data values, you get X dichotomous data with a value of 0 or 1 each.

  Examples:

  - Under the age of 30 vs. Over the age of 30
  - Default on a loan

2.2.2. Structured vs. Unstructured Data

Structured data: Data that conform to a predefined data model.

- They come in a tabular (grid) format; each row is a record (case) and each column is a variable (attribute)
- These rows and columns can, but do not have to, be labeled
- The cells at each intersection of the rows and columns contain values
- They are easy to work with as everything is set up for processing

  Examples:

  - Spreadsheets
  - CSV (Comma Separated Values) files
  - Relational databases

Semi-structured data: Data that do not conform to a tabular data model but nonetheless have some structure.

- The data do not reside in fixed records or fields
- They contain elements (e.g., tags or other markers) that enforce hierarchies of records and fields within the data
- To be found extensively on the web or social media

  Examples:

  - HTML (Hypertext Markup Language) files
  - XML (Extensible Markup Language) files
  - JSON (JavaScript Object Notation, pronounced "Jason") files

Unstructured data: Data that are not organized in a predefined manner. They are not arranged according to a pre-set data model or schema and cannot be stored in the form of rows and columns.

- They come in large collections of files
- They are not easily searchable
- It is difficult to manage, analyze, and protect them with mainstream relational databases
- They require other alternative platforms for storing and processing

  Examples:

  - Text documents (word processing files, pdf files, etc.)
  - Presentations
  - Emails

- Media files (images, audio, video)
- Customer feedback (written and spoken complaints)
- Invoices
- Sensor data (as collected by mobile devices, activity trackers, cars, airplane engines, machines, etc.)

Metadata: These are data about data, providing information about other data.

- Metadata summarize basic information about data
- Metadata can be created (and enriched) automatically and manually
- It concerns various forms of data, e.g., images, videos, emails, social media posts, etc.

  Examples:

  - Author
  - Timestamp (i.e., date and time created)
  - Geo-location
  - File size
  - File type (JPEG, MP3, AVI, etc.)

2.2.3. Data at Rest vs. Data in Motion

Data at rest: Data that are stored physically on computer data storage, such as cloud servers, laptops, hard drives, USB sticks, magnetic tapes, etc.

- They are inactive most of the time, but meant for later use
- Data at rest are vulnerable to theft when physical or logical access is gained to the storage media, e.g., by hacking into the operating system hosting the data or by stealing the device itself

  Examples:

  - Databases
  - Data warehouses
  - Spreadsheets
  - Archives

Data in motion: Data that are flowing through a network of two or more systems or temporarily residing in computer memory.

- They are actively moving from device to device or network to network
- Data in motion are meant to be read or updated
- Data in motion are often sent over public networks (such as the internet) and must be protected against spying attacks

  Examples:
  - Data of a user logging into a website
  - Telemetry data
  - Video streaming
  - Surveillance camera data

2.2.4. Transactional vs. Master Data

When working with data involving larger systems, such as ERP (Enterprise Resource Planning, e.g., S/4HANA) or CRM (Customer Relationship Management, e.g., Salesforce Marketing Cloud), people make a distinction between two other types of data.

Transactional data: Data recorded from transactions. They are volatile, because they change frequently.

Examples:

- Purchase orders
- Sales receipts
- Bank transaction history
- Log records

Master data: Data that describe core entities around which business is conducted. Master data may describe transactions, but they are not transactional in nature. They change infrequently and are more static than transaction data.

Examples:

- Prospects
- Customers
- Accounting items
- Contracts
-

Illustrative example:

If a manufacturer buys multiple pieces of equipment at different times, a transaction record needs to be created for each purchase (transactional data). However, the data about the supplier itself stay the same (master data).

Challenges of master data management:

- Master data are seldom stored in one location; they can be dispersed in various software applications, files (e.g., databases, spreadsheets) or physical media (e.g., paper records)
- Various parts of the business may have different definitions and concepts for the same business entity
- Master data are often shared and used by different business units or corporate functions

2.2.5. Big Data

Big data are data that are too large or too complex to handle by conventional data-processing techniques and software. They are characterized across *three* defining properties:

1.Volume: This aspect refers to the amount of data available. Data become big when they no longer fit on a desktop or laptop RAM.

- RAM (Random Access Memory) can be likened to a computer's short-term memory
- It is fast to read from and write to

- RAM is different from storage capacity, which refers to how much space the hard disk drive (HDD) or solid-state drive (SSD) of a device can store
- The amount of RAM in a device determines how much memory the operating system and open applications can use
- Big data starts with the real-time processing of gigabytes (GB) of data, where 1 GB = 1000 megabytes (MB)- this is equivalent to a document with about 75'000 pages
- Companies can also store terabytes (1 terabyte =1000 gigabytes) or petabytes (1 petabyte = 1000 terabytes) of data

2.Variety: This aspect refers to the multiplicity of types, formats, and sources of data available

- In the past, companies only had structured data at their disposal
- Nowadays, they use data from different sources, e.g. ERP, CRM, Supply Chain Management system; and of different structure
- Data users need to structure and clean the data before they can analyze them, which can take a lot of time

3.Velocity: This aspect refers to the frequency of the incoming data.

- Systems need to log users' activities as data points in order to provide the seamless experience customers expect
- The update needs to be made in a near real time (e.g., daily, several times a day) or even real-time manner

- Big data flow from different sources in real time into a central environment

  Examples:

  - Sensor data produced by autonomous cars (up to 3.6 TB per hour, from about 100 in-built sensors constantly monitoring speed, engine temperature, braking processes, etc.)
  - Search queries on Google or other search engines (40'000 per second, or 3.5 billion per day)
  - Data generated by aircraft engines (1 TB per flight according to GE)

Some sources mention **"Veracity" (or Validity)** as a fourth defining property. This aspect refers to the accuracy or truthfulness of a data set.

- It cannot be used to describe how "big" data are
- It is equally important for "small" and "big" data

2.3. Storing Data

There are many applications and tools needed for end-to-end data management.

Examples:

- Data security management
- Master data management
- Data quality management
- Metadata management

The end consumers of data can access data through databases, data warehouses, data lakes, etc. Depending on the type of data stored and processed, a distinction is made between “traditional” data systems and “big” data systems.

2.3.1. Database

A **database** is *a systematic collection of data*. A computer system stores and retrieves the data electronically.

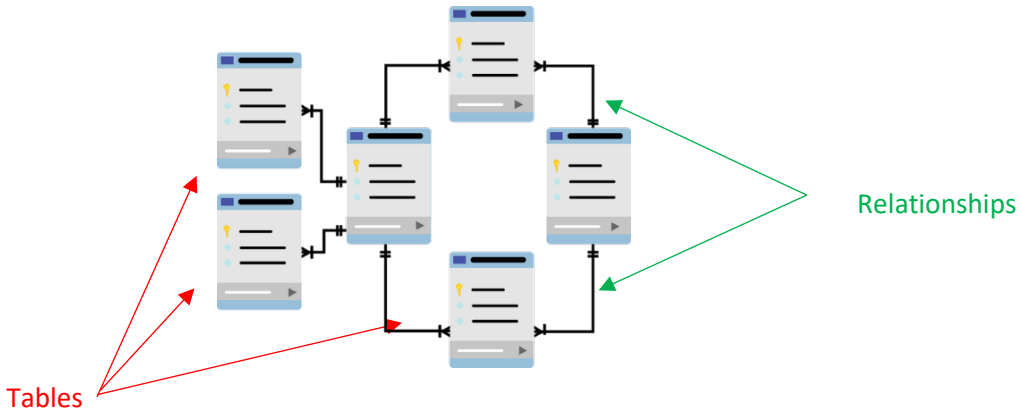
Simple databases are made of tables, which consist of rows and columns:

- Rows - represent a set of related data and has the same structure (i.e customer record)
- Columns (a.k.a attribute, field) - made of data values (text values or numbers) describing the item in the row. It provides one single value for each row

Example: customer ID, customer name, customer address

- A data model shows the interrelationships and data flow between different tables. It documents the way data are stored and retrieved

Picture of a data model:



Relational databases:

- Data are organized into tables that are linked (or "related") to one another
- Each row in the table contains a record with a unique ID (the key), e.g., customer ID
- Relational databases are very efficient and flexible when accessing structured information

SQL: One term that you will often hear in the context of relational databases is SQL.

- SQL stands for Structured Query Language
- It is a programming language designed to facilitate the retrieval of specific information from databases
- It is used to query, but also to manipulate and define data, and to provide access control
- It is easy to learn, also for non-technical users
- Common database software products include Microsoft SQL Server or MySQL

2.3.2. Data Warehouse

A **data warehouse** is a central repository of integrated data from one or more disparate sources.

- It is designed for query and analysis
- It contains large amounts of historical data

- It allows integrating data from applications and systems in a well-architected fashion
- It is built to serve and contain data for the entire business
- Access to it is strictly controlled
- It is difficult to query the data needed in a data warehouse
- Data warehouse constitutes a core component of business intelligence (BI)

Business intelligence (BI) includes:

- Reports (sent automatically to relevant users)
- Dashboards
- Visual exploration of data
- Multidimensional queries (e.g., for the analysis of sales across products, geography, and time)
- Slicing and dicing of data (e.g., to look at them in various ways)
- Analysis of geographic and spatial data (e.g., using maps)

2.3.3. Data Marts

A data mart is a subset of a data warehouse focused on a single functional area, line of business, department of an organization. IBM, Oracle, SAP or Teradata use data warehouses or data marts.

- They are designed for the needs of specific teams or departments
- These users usually control their respective data mart
- Data marts make specific data available to a specific set of users

- Enables users to quickly access critical information and insights without having to search through a central data archive

Data marts vs. data warehouses - advantages:

- Quick, easy, and efficient (cheap) access to data
- Faster deployment due to fewer sources involved and comparatively simpler setup
- Flexibility because of smaller size)
- Better performance due to faster data processing
- Easier maintenance when the relevant department is under control
- Departments within the organization do not interfere with each other's data

2.3.4. The ETL Process

Building data warehouses and data mart involves copying data from one or more source systems into a destination system.

The data in the destination system may represent the data differently from the source, involving a series of treatments of the data.

To prepare data and deliver these in a format so that they can feed into or be read by applications, companies go through an ETL process.

ETL Process

*E*xtracting (the data from source)

*T*ransforming (the data while in transit)

*L*oading (the data into the specified target data warehouse or data mart)

Transforming includes:

- Selecting specific columns
- Recoding value (e.g., No -> 0, Yes -> 1)
- Deriving new value through calculation
- Deduplicating, i.e., identifying and removing duplicate, records
- Joining data from multiple sources
- Aggregating multiple rows of data
- Conforming data so that separate sources can be used together
- Cleansing data to ensure data quality and consistency
- Changing the format and structure so that the data can be queried or visualized

Companies offering ETL tools include Informatica, IBM, Microsoft and Oracle

2.3.5. Apache Hadoop

Data storage in the past

A company needed data to:

- Get a snapshot of its financial performance

- Determine sales trends in different markets
- Find out what customers thought of their products

Data was "regular", i.e., there were limited volumes of structured data at rest.

Data storage today

Companies store big data "just in case". For this, they need a broader range of business intelligence and analytical capabilities, which should enable them to:

- Collect data of unlimited volume (in depth and breadth)
- Process structured as well as semi-structured and unstructured data
- Support real-time data processing and analysis
- All these data cannot all be fit in a single file, database or even a single computer
- Processing them simultaneously is also practically impossible with one computer

Apache Hadoop was the first major solution to address these issues, becoming synonymous with "big" data.

Definition:

"*Hadoop is a set of software utilities designed to enable the use of a network of computers to solve "big" data problems, i.e., problems involving massive amounts of data and computation."*

This technology allows you to:

- Spread the same data set across a large multitude of computers

- Make data available to users far faster than with traditional data warehousing
- Stream data from different sources and in different formats with Hadoop Distributed File System (HDFS), ignoring the rules and restrictions imposed by a relational database

2.3.6. Data Lake

Definition:

"A data lake is a single repository of data stored in their native format.". They can store structured, semi-structured and unstructured data, at any scale.

Examples:

- Source system data (e.g., from ERP, CRM)
- Sensor data (e.g., from machines, smart devices)
- Text documents
- Images
- Social media data

- Data lakes are usually configured on a cluster of cheap and scalable commodity hardware.
- They constitute a comparatively cost-effective solution for storing large amounts of data
- Companies commonly dump their data in the lake in case they need them later
- Data lakes are advantageous in scenarios where it is not yet clear whether the data will be needed at all

- The lack of structure of the data make them more suitable for exploration than for operational purposes

Data lakes vs. data warehouses – differences:

- Data lakes contain information that has not been pre-processed for analysis
- Data lakes retain all data (not just data that can be used today)
- Data lakes store data for no one particular purpose
- Data lakes support all types of data (not just structured data)
- Data lakes and data warehouses complement each other. Companies often have both

Data lakes – risks:

- Data lakes need maintenance
- If this does not take place, data lakes can deteriorate or become inaccessible to their intended users; such valueless data lakes are also known as data swamps

Vendors that provide data lake technology: Amazon, Databricks, Delta Lake, or Snowflake

2.3.7. Cloud Systems

Data lakes can be stored in two locations – "*on premise*" and "*in the cloud*".

On premise: Data lakes are stored within an organization's data centers. Advantages:

- Security: The owner's data are under control within its firewall
- Speed: Internet tends to be faster within the same building

- Cost: It is cheaper purchase one's own hardware than to lease it from a third-party service provider (at least if the system is under constant use)

In the cloud: Data lakes are stored, using Internet-based storage services such as Amazon Web Services (AWS), Google Cloud, Microsoft Azure, etc. Advantages:

- Accessibility: Since the data are online, users can retrieve them wherever and whenever they need them
- Adaptability: Additional storage space and computing power can be added immediately as needed
- Convenience: All maintenance (e.g., replacement of broken computers) is taken care of by the cloud service provider
- Resilience: Service providers typically offer redundancy across their own data centers by making duplicates of their clients' data, retained as a fallback

2.3.8. Edge Computing

Edge (Fog) computing: The computing structure is located between the cloud and the devices that produce and act on the data.

- The "fog" is like a cloud, expect that is it on the ground
- Fog computing allows to bring the cloud down to the users
- These devices, while connected to the central storage and processing unit, are located at the "edge" of the network

- When an IoT device generates data, these can be processed and analyzed on the device itself without having to be transferred all the way back to the cloud
- It can also take on some of the workload from the central computer
- Examples of such devices: mobile phone, smart watch, video cameras, sensors, industrial controllers, routers, etc.

Benefits:

- Lower latency: smaller delays in response times increasing the processing speed

Example:

Autonomous cars can recognize obstacles on the road in real-time

- Lower dependency on connectivity: a connection to the internet is not required to get the information or to do the processing

Example:

A fitness tracker delivers performance statistics even when the user is offline

- Privacy: Data are not exposed to the internet or to the central storage or processing unit

Example:

During the COVID-19 pandemic, when contact tracing apps were launched. Many countries opted for a version where the storage and processing

of users' geo-location data take place on their mobile devices (as opposed to the app operator's central server).

2.3.9. Batch vs. Stream Processing

Batch processing: Data are processed in large volumes all at once.

- A batch is a block of data points collected within a given period (e.g., a day, a week) and can easily consist of millions of records
- After the collection is complete, the data are then fed into a system for further processing
- Batch processing is efficient for large sets data at and where a deeper or complex analysis of the data is required
- Consumers can explore and use the data to develop statistical models, e.g., for predictions
- This approach should be chosen when it is more important to crunch large volumes of data

Example: A fast-food chain keeps track of daily revenue across all restaurants. Instead of processing all transactions in real-time, it aggregates the revenue and processes the batches of each outlet's numbers once per day.

Stream processing: Data (in motion) are fed into a system as soon as they are generated, one bit at a time.

- A better choice when the speed matters

- Each data point or "micro-batch" flows directly into an analytics platform, producing key insights in near real-time or real-time.
- Good for supporting non-stop data flows, enabling instantaneous reactions
- It works best when speed and agility are required
- Common fields of applications include:
  - Cybersecurity
  - Stock trading
  - Programmatic buying (in advertising)
  - Fraud detection
  - Air traffic control
  - Production line monitoring

Example: A credit card provider receives data related to a purchase transaction as soon as the user swipes the card. Its systems can immediately recognize and block anomalous transactions, prompting additional inspection procedures. In case of non-fraudulent charges are approved without delay, so that customers do not have to wait unnecessarily.

Both processing techniques:

- Require different types of data (at rest vs. in motion)
- Rely on different infrastructure (storage systems, database, programming languages)
- Impose different processing methods
- Involve different analytics techniques

- Address different issues
- Serve different goals

2.3.10. Graph Database

Traditional relational database systems are not equipped to support connections across beyond a certain degree. To achieve that, companies use graph databases.

Graph (semantic) databases:

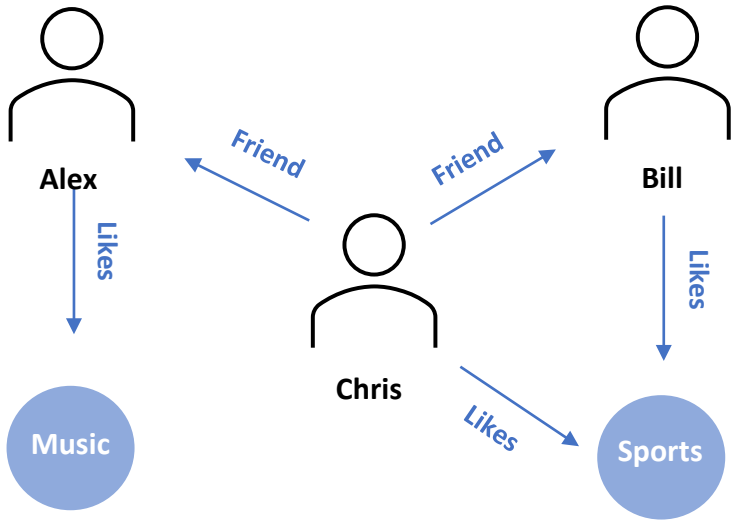
- Store connections alongside the data in the model
- Treat the relationships between data as equally important to the data themselves
- Show how each individual entity connects with or is related to others
- The networks of connections that can be stored, processed, modified, queried, etc. are known as graphs

A graph consists of **nodes** and **edges**:

- A node can have any number of properties or attributes, e.g., age, gender, nationality
- An edge represents the connections between two nodes
- Edges describe relationships between entities, e.g., friendship, business association, ownership, action, preference
- There is no limit to the number and kind of relationships a node can have
- An edge always has a start node, end node, type, and direction

- Like nodes, edges can have properties, which are usually quantitative, e.g., weights, strengths, distances, ratings, time intervals, costs

Example of a simple graph:



The people (Alex, Bill, Chris) and hobbies (Music, Sports) are data entities, represented by “nodes”.

Example: On Facebook, the nodes represent users and content, while the edges constitute activities such as “is a friend”, “posted”, “like”, “clicked”, etc.

Business use cases:

- Marketing, e.g., determine "friends of friends" and identify common interests of customers on social media
- Recommendation engines, e.g., based on the logic "customers who bought this also looked at..."

- Fraud detection through the identification of clusters or people of events that are connected in unusual ways

Technologies or company names related to graph databases: OrientDB, ArangoDB, neo4j, or TigerGraph.

3. Using Data

3.1. Analysis vs. Analytics

Definition:

"*Analysis is a detailed examination of anything complex in order to understand its nature or to determine its essential features.*"

- Data analysis is the in-depth study of all the components of a given data set
- Analysis is about looking backward to understand the reasons behind a phenomenon
- It involves the dissection of a data set and the examination of all parts individually and their relationship between one another
- The ultimate purpose of analysis is to extract useful information from the data (discovery of trends, patterns, or anomalies)
- It involves the detailed review of current or historical facts
- The data being analyzed describe things that already happened in the past

Examples:

- Comparison of the sales performances across regions or products
- Measurement of the effectiveness of marketing campaigns
- Assessment of risks (in finance, medicine, etc.)

Definition:

"Analytics is a broader term covering the complete management of data."

- It encompasses not only the examination of data, but also their collection, organization, and storage, as well as the methods and tools employed
- Data analytics implies a more systematic and scientific approach to working with data throughout their life cycle, which includes: data acquisition, data filtering, data extraction, data aggregation, data validation, data cleansing, data analysis, data visualization, etc.
- Analytics allows to make predictions about the future or about objects or events not covered in the data set

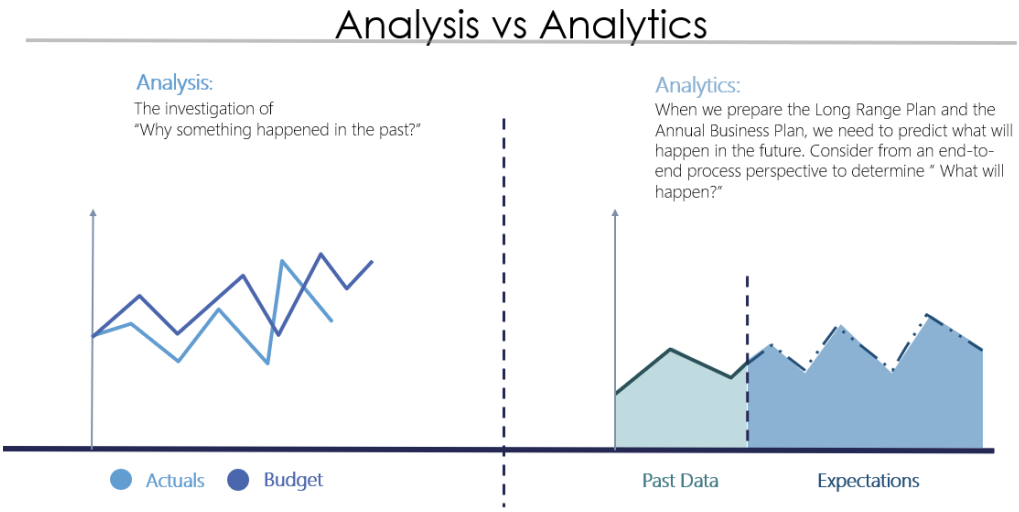
Examples:

- The semi-automatic or automatic analysis of data
- The extraction of previously unknown patterns
- The grouping of objects
- The detection of anomalies
- The discovery of dependence

Analysis vs. Analytics – differences:

- Analysis only looks at the past, whereas analytics also tries to predict the future
- Analysis can be performed with any spreadsheet software, in particular Microsoft Excel.

- For Analytics, you need an analytics tool, such as Python



3.2. Statistics

Definition:

"Statistics concerns the collection, organization, analysis, interpretation, and presentation of data."

"*Population is the entire pool from which a statistical sample is drawn.*"

Well-known software tools for statistical analysis: IBM SPSS Statistics, SAS, STATA, and R.

Types of statistics:

- Descriptive statistics
- Inferential statistics

3.2.a Descriptive Statistics:

- Summarizes (describes) a collection of data
- Reveals useful or at least interesting things about the data, in particular "what is going in there".
- Presents and communicates the results of experiments and analyses.

Examples of questions to be answered through descriptive statistics:

- How many people in Georgia voted for Joe Biden during the US Presidential Election in 2020?
- What is the proportion of favorable opinions for Brand X among adults aged 18 to 34?
- What is the average salary of brand managers?

3.2.b Inferential Statistics:

- Helps us draw conclusions (inferences) about things that we do not fully grasp
- Helps us understand more about populations, phenomena, and systems from sample data (induction)
- Enables professionals to predict what is likely to happen if we make a specific decision or act on the entity in this or that manner

Examples of questions to be answered through inferential statistics:

- Is there a correlation between age and customer satisfaction?
- Does this new packaging significantly increase the sales volume of this product?
- In this country, is the brand preference of women the same as that of men?

Computational statistics (statistical computing):

- A combination between computer science and statistics
- The focus lies on computer-intensive statistical methods, e.g., a very large sample size

Practical applications of statistical computing:

- Econometrics
- Operations research
- Monte Carlo simulation

3.3. Business Intelligence (BI)

Definition:

"Business intelligence (BI) comprises the strategies and technologies used by enterprises for the data analysis of business information."

Practical Applications:

- Operating decisions, e.g., the allocation of marketing budget to different campaigns, brand positioning, pricing
- Strategic decisions, e.g., international expansion, opening or closing of plants, priority setting
- Enterprise reporting – the regular creation and distribution of reports describing business performance
- Dashboarding – the provision of displays that visually track key performance indicators (KPIs) relevant to a particular objective or business process
- Online analytical processing (OLAP) – querying information from multiple database systems at the same time, e.g., the multi-dimensional analysis of data
- Financial Planning and Analysis – a set of activities that support an organization's financial health, e.g., financial planning, forecasting, and budgeting

Vendors providing visualization or business intelligence software: MicroStrategy, Microsoft (with Power BI), Qlik, or Tableau.

3.4. Artificial Intelligence (AI)

Definition:

"*Artificial intelligence (AI) is the ability of a computer, or a robot controlled by a computer, to do tasks that are usually done by humans because they require human intelligence and discernment.*"

Related Fields:

Artificial intelligence research relies on approaches and methods from various fields: Mathematics, statistics, economics, probability, computer science, linguistics, psychology, philosophy, etc.

Practical Applications:

- Recognizing and classifying objects (on a photo or in reality)
- Playing a game of chess or poker
- Driving a car or plane
- Recognizing and understanding human speech
- Translating languages
- Moving and manipulating objects
- Solving complex business problems and making decisions like humans

Examples:

- In 2016, AlphaGo (developed by DeepMind Technologies, later acquired by Google) defeated Go champion LEE Sedol
- In 2017, Libratus (developed at Carnegie Mellon University) won a Texas hold 'em tournament involving four top-class human poker players
- In 2019, a report pooling the results of 14 separate studies revealed that AI systems correctly detected a disease state 87% of the time (compared with 86% for human healthcare professionals) and correctly gave the all-clear 93% of the time (compared with 91% for human experts)
- In 2020, researchers of Karlsruhe Institute of Technology developed a system that outperforms humans in recognizing spontaneously spoken language with minimum latency.

Narrow AI:

- AI programs that are able to solve one single kind of problem
- Narrow AI applications that work in different individual domains could be incorporated into a single machine
- Such a machine would have the capacity to learn any intellectual task that a human being can, a.k.a. “artificial general intelligence” (AGI)

Conclusions: AI is used to improve the effectiveness or efficiency of processes or decisions. It should be thought of as a facilitator of human productivity – not as a replacement for human intelligence.

3.5. Machine Learning (ML)

Definition:

"Machine learning (ML) is the study of computer algorithms that improve automatically through experience."

Machine learning is about teaching computers to:

- Find hidden insights without being explicitly programmed where to look
- Make predictions based on data and finding
- Produce a "model", e.g., a predictive model

A model can be described as:

- A formula that takes some values as input and delivers one or more values as output
- The computer can use this description to learn and then make predictions
- It is only a mathematical description of phenomena or events, and can never fully represent the reality

The creation of a model requires two elements: 1) an algorithm and 2) data

Definition:

"An algorithm is a procedure for solving a mathematical problem in a finite number of steps that frequently involves repetition of an operation."

An algorithm:

- Provides step-by-step guidance on what to do to solve the problem

- Can be run again with a few parameters variations, to see if the new configuration leads to a better result
- It can take several iterations for the algorithm to produce a good enough solution to the problem
- Selects the model that yields the best solution
- Machine learning algorithms find their way to better solutions without being explicitly programmed where to look
- Users must determine how to define when the problem is solved or what kind of changes need to be made for each iteration

Training data:

- These are past observations, used to "feed" the algorithm so that it can gain initial experience
- These observations represent units of information that teach the machine trends and similarities derived from the data
- The machine gets better with every iteration
- Once the algorithm is able to distinguish patterns, you can make predictions on new data
- The process an algorithm goes through training data again and again is called "training the model".

3.6. Supervised Learning

Definition:

"Mapping is the association of all elements of a given data set with the elements of a second set."

Supervised learning splits the data into:

- Training data: an algorithm analyzes the training data (related to existing observations)
- Validation data: an algorithm produces a function for the mapping of validation data

During the training process:

- The algorithm "learns" the mapping function from the input to the output
- The algorithm learns by comparing its own computed output with the correct outputs (provided in the training data) to find errors
- The goal is to plot a function that best approximates the relationship between the input and output observable in the training data
- Once the patterns and relationships have been computed, these can then be applied to new data, e.g., to predict the output for new observations
- The process is called "supervised" learning, because it is as if the learning took place in the presence of a supervisor checking the outcome
- In every iteration, the model provides a result

- That result is then compared with the "correct" answer (provided by the labels in the training data)
- Direct feedback is given to the algorithm, which takes it into consideration for its next iteration
- It requires technical proficiency to develop, calibrate, and validate supervised learning models
- The models produced through supervised learning are such that they are fed an input and they deliver an output
- Each observation in the training data set is tagged with the outcome the model is supposed to predict
- Outcome can also be continuous or categorical with two or more classes

Types of supervised machine learning techniques:

- Regression (when the output is continuous data)
- Time series forecasting (when input and output are arranged as a sequence of time data)
- Classification (when the output is categorical data)

3.6.1.Regression

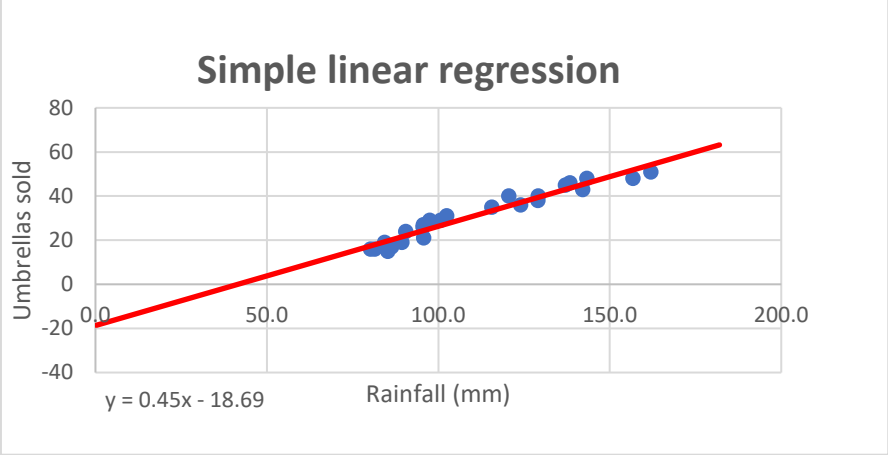
Regression is a method that is used to determine the strength and character between a dependent variable and one or more independent variables.

- The dependent variable is what is to be explained, or what is to be predicted

- The independent variables (a.k.a predictors) are the explanatory variables, i.e., the variables that are used to make the prediction
- Regression is intended for the prediction of continuous variables, i.e. numeric data that have an infinite number of values in a range between any two values
- The goal of the regression analysis is to draw a red line on a plot, which is the visualization of an equation connecting both variables
- Regression is particularly useful when the independent variables represent factors that can be controlled (which is not the case for rainfall).

Simple linear regression - we can make predictions as follows "given a particular x value, what is the expected value of the y variable".

Example:

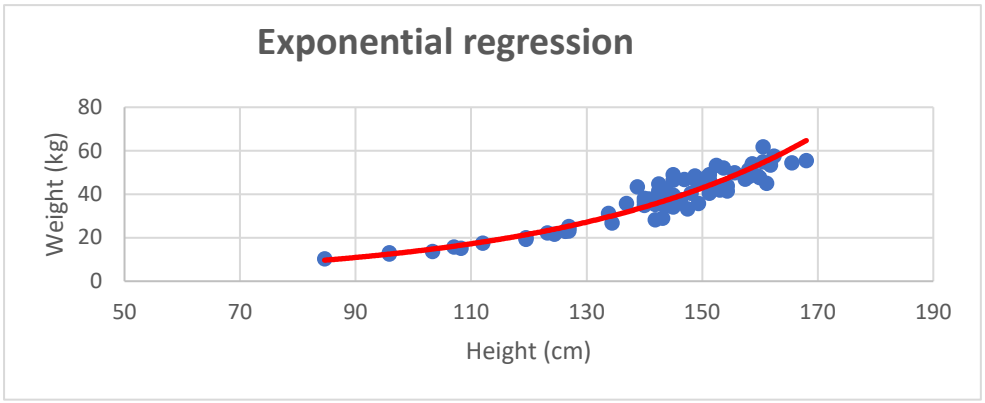


- The dots on the scatter plot correspond to the observed data, each of them representing one month (24 months in total)

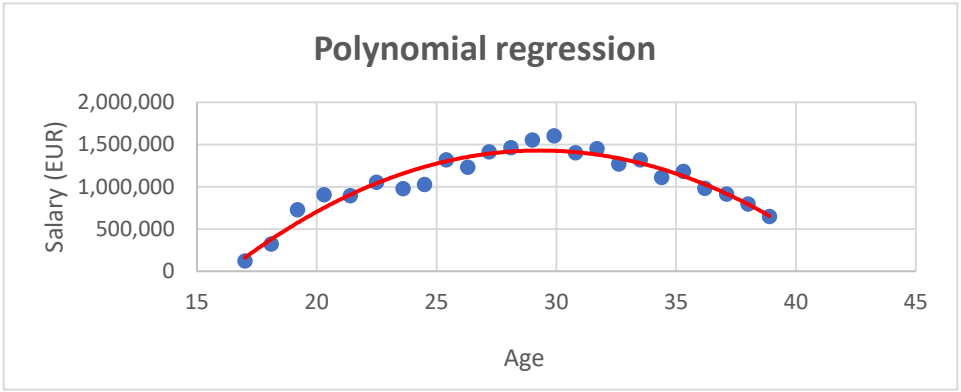
- For each month, we can see the rainfall in mm (the independent variable, on the horizontal axis) and the number of umbrellas sold (the dependent variable, on the vertical axis)
- There's a positive relationship between both variables: the larger the rainfall, the higher the number of umbrellas sold – and vice versa

Polynomial regression - when the relationship between the dependent and independent variable is non-linear.

Example 1:



Example 2 :



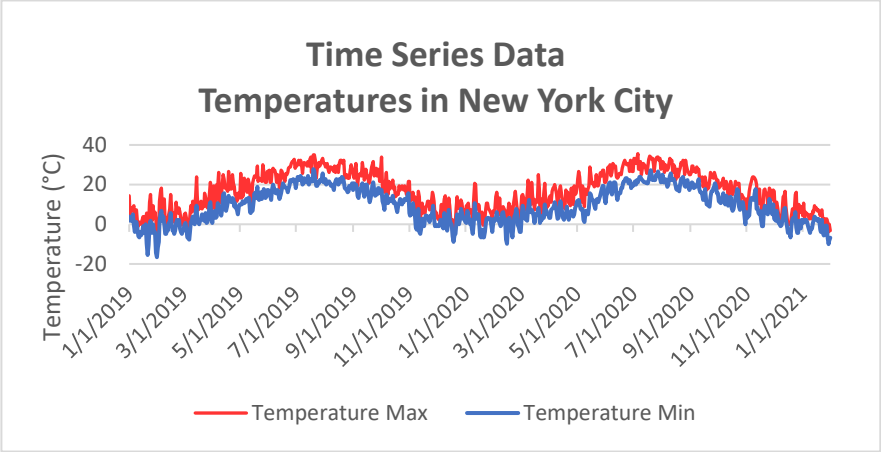
Multivariate regression - when we use two or more independent variables to predict a single outcome. The number of independent variables is virtually unlimitedMultivariate regression can determine how the variables relate to one another and what variables matter the most when explaining a phenomenon:

3.6.2. Time Series Forecasting

Definition:

"Time series forecasting is defined as the use of models to predict future values based on previously observed values."

- It is a supervised machine learning, where past values can be the training data used for the mapping of future values
- A time series is a sequence of discrete-time data, taken at equally spaced time intervals, e.g., days, months, or years
- We employ previous time periods as input variables, while the next time period is the output variable
- Time series analysis has the objective to forecast the future values of a series based on the history of that series
- Time series are used in various fields, e.g., economics, finance, weather forecasting, astronomy, supply chain management, etc.
- Time series are typically plotted via temporal line charts (run charts):



Examples:

- Daily closing prices of stocks
- Weekly sales volume of laundry detergents
- Monthly count of a website's active users
- Annual birth or death rate for a given country

3.6.3. Classification

Definition:

"*Classification is a supervised machine learning approach where a class (also called category, target, or label) is predicted for a new observation. The main goal is to identify which class the new observation will fall into. "*

- The learning is based on a training set of data with observations for which category membership is already known
- Classification can be performed on both structured and unstructured data
- The output and input data must be categorical
- Classification is best use when the output has finite and discrete values
- When there are only two categories, we have "binary" classification, e.g. Yes vs. No, Churn vs. Non-churn, etc.
- For more categories, the term used is "multi-class" classification, e.g., Positive/Neutral/Negative, Win/Draw/Loss, 0 to 9

Email spam detection - example:

Spam detection can be considered as a binary classification problem, since there two classes – "Spam" or "Not spam", respectively Spam "yes" or "no". The input data consist of variables, such as:

- Title includes keywords such as "winner", "free", "dollar", etc.
- Email body contains special formatting (such as bold, entire words in capital letters, etc.)
- Number of addressees in the "To" field of the email
- Email includes one or more attachments

The model (here - a mapping function) quantifies the extent to which the input variables affect the output (or "predicted") variable. The objective is to approximate the mapping function so that when we can predict the nature of an email (spam or not) as soon as we get an email in the mailbox.

Classifier :

An algorithm that implements classification. Classifiers are used for all kinds of applications:

- Image classification
- Fraud detection -> Is this transaction fraudulent?
- Direct marketing -> Will this customer accept this offer?
- Churn prediction -> What is the chance that this subscriber switch to another provider (high, medium, low)?
- Credit scoring -> Will this prospective borrower default?

3.7. Unsupervised Learning

Definition:

"An unsupervised learning is based on an algorithm that analyzes the data and automatically tries to find hidden patterns."

- No guidance or supervision required
- No specific desired outcome or correct answer is provided
- The only objective is for the algorithm to arrange the data as best as it can
- It identifies structure in the data
- The training data are unlabeled; more cost-effective than supervised techniques
- Here, analysts do not worry about quality and accurate labeling
- In unsupervised learning, only input data are provided, but no corresponding output
- No explicit instructions on what to do with the data is given
- The algorithm does not know what is right or wrong
- It is difficult to assess or compare the truthfulness of results obtained
- The two most useful unsupervised learning techniques are clustering and association

Unsupervised learning techniques are used to:

- Detect anomalies
- Fraud detection (via flagging of outliers)

- Predictive maintenance (via discovery of defective parts in a machine or system)
- Network intrusion detection

- Reduce features in a data set
- Describing customers with 5 attributes almost as precisely as with 10 attributes

3.7.1. Clustering Analysis

Definition:

"Clustering is the splitting of a data set into a number of categories (or classes, labels), which are initially unknown."

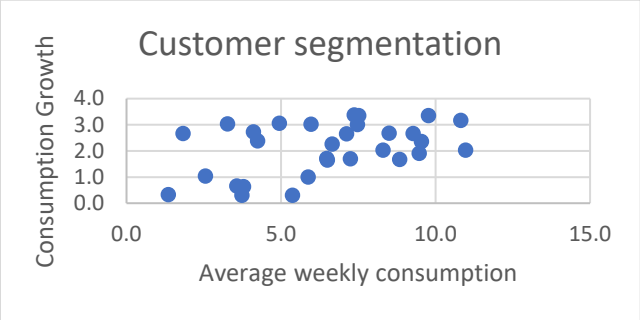
- You do not know in advance what we are looking for
- The data are unlabeled
- The objective is to discover possible labels automatically
- These categories produced are only interpreted after their creation
- A clustering algorithm automatically identifies categories to arrange the data in a meaningful way
- The grouping of the data into categories is based on some measure of inherent similarity or distance
- The algorithm must group the data in a way that maximizes the homogeneity within and the heterogeneity between the clusters
- The interpretation of the clusters constitutes a test for the quality of the clustering

Example:

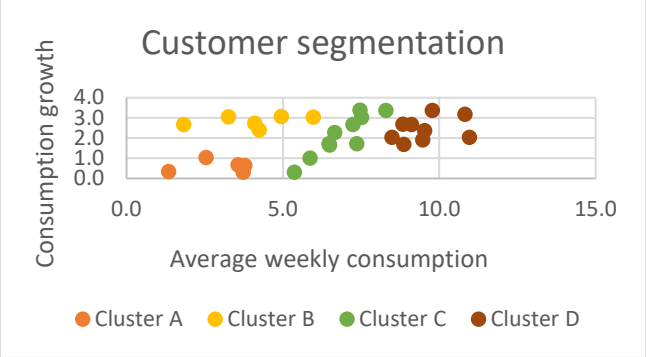
30 data points (each data point corresponding to one consumer of chocolate bars) and two variables:

X-axis - the average weekly consumption of chocolate bars in the last 12 months

Y-axis - the growth in consumption of chocolate bars in the 12 months, compared to the previous 12 months



By applying the clustering technique, 4 distinct categories emerge:



These four customer segments could be specifically targeted with promotions or adverts. Cluster B could be lured with discounts; Cluster A might prefer healthier products.

How to interpret these groups is a question to be answered by the business:

- What do you know about chocolate bar consumption habits and their effects on buying patterns?

- Why did these patterns emerge and how are they relevant to our business?
- What could we offer to consumers in Cluster C (average weekly consumption) so that they buy more of our products?

Practical applications:

- Customer segmentation
- Sorting of e-mails or documents into themes
- Grouping of retail products based on customer behavior or other characteristics
- Image recognition, e.g., grouping of similar images

3.7.2. Association Rules

Definition:

"Association rule learning is *an unsupervised machine learning method to discover interesting associations (relationships, dependencies) hidden in large data sets."*

The associations are usually represented in the form of rules or frequent item sets.

Practical applications:

- Market basket analysis: The exploration of customer buying patterns by finding associations between items that shoppers put into their baskets, e.g., "tent and sleeping bags" or "beer and diapers"; it produces the probability of a joint occurrence

- Product recommendation: When a customer puts a tent in his online shopping cart, the website displays an offer for a sleeping bag
- Design of promotions: When the retailer offers a special price for just one of the products (crackers) but not for the other one at the same time (dip)
- Customization of store layout: When you increase customer traffic so that customers have more opportunities to buy products; strawberries and whipped cream can be placed at the opposite ends of the store

3.8. Reinforcement Learning

Definition:

"Reinforcement learning is a technique where *a machine ("agent") learns through trial and error, using feedback from its own actions."*

- The learning is based on the rewarding of desired behaviors and/or the punishing undesired ones ("the carrot and stick for machines")
- The agent is given an overarching goal to reach. It performs tasks randomly, and "sees what happens" in every iteration. Over time, it learns what brings a reward or a punishment
- The agent is programmed to seek the maximum overall reward when interacting within a given situation ("environment")
- Used for the development of robots, autonomous vehicles, and game-playing programs, and in various business areas, e.g., marketing

Reinforcement learning vs. supervised learning:

- In both approaches, algorithms are given specified goals
- They use mapping between input and output; outcomes can be positive or negative
- Reinforcement learning does not require labelled data sets
- The agent is not given any directions on how to complete the task
- The agent can uncover entirely new solutions

Example:

In 2016, AlphaGo defeated LEE Sedol, the world champion in the game of Go. Within a period of just a few days, AlphaGo had accumulated thousands of years of human knowledge, enough to beat a champion with 21 years of professional experience.

3.9. Deep Learning

Definition:

"Deep learning is a subfield of machine learning, applying methods inspired by the structure and function of the human brain."

- Deep learning can be supervised, unsupervised or take place through reinforcement learning
- With deep learning, computers learn how to solve particularly complex tasks – close to what we describe as "artificial intelligence"
- Deep learning teaches machines to imitate the way humans gain knowledge

- The algorithm performs a task repeatedly, adjusting it a bit to improve the outcome
- Neural networks can be taught to classify data and identify patterns like humans
- When the machine receives new information, the algorithm compares it with known objects and tries to draw similar conclusions
- The more it knows about the objects' distinguishing characteristics, the more likely the algorithm is to solve the problem, e.g., recognize a certain animal on an image
- The system improves its ability as it gets exposed to more examples

Artificial neural networks (ANN) - The algorithms used to achieve deep learning.

- ANN were inspired by the way the brain processes information and how communication nodes are distributed
- A neural network is made of a collection of connected nodes (called neurons)
- By connecting the nodes to other another, information can "flow" through the network
- Neurons are organized in layers
- The more layers involved, the "deeper" the neural network

Practical applications:

- Image recognition
- Autonomous driving: The computer can differentiate between pedestrians of different sizes, e.g., children vs. adults
- Speech recognition, e.g., to detect "Hey Siri" spoken by iPhone users, language translation, e.g., Google Translate
- Product recommendations, e.g., Netflix or YouTube

3.10. Natural Language Processing (NLP)

Definition:

"Natural Language Processing (NLP) is a field of artificial intelligence concerned with the interactions between computers and human (natural) language."

- It enables computers to process and analyze large amounts of natural language data
- The goal of NLP is to decipher and to derive meaning from human languages
- It taps into the treasure troves of unstructured data that are held by companies (emails, chats, blogs, images, etc.)
- There are two sub-topics of NLP

1. Natural Language Understanding (NLU) – It deals with the reading comprehension of machines, i.e., the ability to process documents, understand its meaning, and to integrate with what the reader already knows.

Sentiment Analysis – the interpretation and classification of emotions in text.

- A machine tries to identify and categorize terms and opinions expressed in a written or spoken text
- The objective is to determine the author's attitude towards a subject
- The output is a sentiment score, like positive, neutral, or negative
- Used in marketing, customer relationship management or customer service, e.g., to determine how they feel about their brands, products, services, etc.
- Brands can apply it to emails, social media posts, phone calls, reviews on eCommerce platforms, etc.
- It can be a valuable source of insights, as the feedback is left in an unprompted fashion
- It delivers information about customers' preferences and choices, and decisions
- The sentiment score is numerical
- It is used in subsequent analyses, e.g., to establish a quantitative relationship between customer satisfaction and revenue

2.Natural Language Generation (NLG) – It deals with the creation of meaningful sentences in the form of natural language.

- It is about transforming structured data into natural language
- An NLG system makes decisions about how to turn a concept into words and sentences
- Less complex than NLU since the concepts to articulate are usually known

- An NLG system must simply choose the right expressions several potential ones
- Classic examples: the production of (written out) weather forecasts from weather data, automated journalism, generate product descriptions for eCommerce sites, interact with customers via chatbots, etc.

4. Reading Data

In some areas of life, the right questions are more important than the right answers. This also applies to working with data. It will often be the case that the ultimately beneficiary of data (for example, the decision makers) is different from the person who processed these data and produced the results (or the answers).

- Although it is important to be able to trust one's own employees and partners, it never harms to get a sound understanding about what was done with these data, i.e., how these were collected, manipulated, cleaned, transformed, analyzed, and turned into insights
- One needs to be able to "read" the data, which includes the assessment of their quality and their description in statistical terms

4.1. Data quality assessment

Data quality is the *state of the information at hand*

Acceptable (or high) data quality - if the data are fit for their intended uses (e.g., planning, decision making, insights generation, statistical analysis, machine learning, etc.)

Poor data quality (e.g., if the data contain errors) - decision makers could be misled or algorithms will not learn efficiently, or perhaps even worse, learn the wrong things.

* "Garbage in, garbage out" (**GIGO**) - Without meaningful data, you cannot get meaningful results

The impact that poor data quality can be significant. Financial expenses, productivity losses, missed opportunities and reputational damage are but some of the costs that poor data quality carry.

**In 2016, IBM estimated that businesses were losing, in the US alone, $3.1 trillion every year due to poor data quality.*

Real examples of blunders (what can go wrong with data):

- A customer makes a spelling mistake when writing his address in the paper registration form
- A clerk misreads the handwriting of the customer and fills in the wrong address into the system
- The CRM system only has space for 40 characters in the "address" field and cuts every character beyond this limit

- The migration tool cannot read special characters, using placeholders instead
- A data analyst makes a mistake when joining two data sets, leading to a mismatch of attributes for all records

Key data flaws:

Incomplete data: This is when there are missing values in the data set.

- If there are too many missing values for one attribute, the latter becomes useless and needs to be discarded from the analysis
- If only a small percentage is missing, then we can eliminate the records with the missing values or make an assumption about what the values could be
- Business decision makers should ask the following questions to the people who processed the data:
- What is the proportion of missing values for each attribute?
- How were the missing values replaced?

Inaccurate data: This can happen in many ways, for examples:

- Spelling mistakes
- Inconsistent units
- A famous case is that of NASA's Mars Climate Orbiter, which in 1999 was unintentionally destroyed during a mission in 1999. The loss was due to a piece of software that used the wrong unit of impulse (pound-force

seconds instead of the SI units of newton-seconds). The total cost of the mistake was estimated at $327.6 million

- Inconsistent formats
- E.g., when numbers are declared as text, which makes them unreadable for some programs
- Impossible values
  - A negative age (for a person)
  - A height of 5.27 meters for a human
  - A household size of 4.8 (for one family)
  - A future birthday (for someone already born)
  - An employer called “erufjdskdfnd”
  - “999” years as a contract duration
- Unusually high or low value does not necessarily have to be inaccurate, but could also be an outlier (i.e., a data point that differs significantly from other observations)

4.2. Data description

When communicating about or with data, it is essential to be able to describe these.

- Although it is not always possible to go through them one by one, there are ways to explore data and to get an overview about what is available.
- The objective is to single out possible issues or patterns that might be worth digging further into
- A description of the data also helps determine how useful they can be to answer the questions one is interested in

Data quality (including completeness and accuracy) are two key properties about the data that should be clarified

Beyond that, the questions that a data consumer should ask are the following:

- What do the data describe?
- Do the data cover the problems we are trying to solve?
- What does each record represent? How granular (vs. coarse) are they?
- How "fresh" are the data? Do they reflect the current situation? When were they collected?

Descriptive statistics can be used to describe and understand the basic characteristics of a data set. A descriptive statistic is a number that provides a simple summary about the sample and the measures. It allows to simplify large amounts of data in a convenient way

Descriptive statistics are broken into two basic categories:

1) Measures of central tendency

2) Measures of spread

4.3. Measures of central tendency

Measures of central tendency (or "measures of center") - focus on *the average or middle values of a data set*

They describe the most common patterns of the analyzed data set. There are three main such measures: 1) the mean, 2) the median, and 3) the mode

The mean is the average of the numbers.

- It is easy to calculate
- It is the sum of the values divided by the number of values

Example: The following data set indicates the number of chocolate bars consumed by 11 individuals in the previous week. The series has 11 values (one value per person):

0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16

Mean = (0 + 1 + 2 + 2 + 3 + 4 + 5 + 6 + 7 + 9 + 16) / 11 = 55 / 11 = 5

This means that any given person in this sample ate 5 chocolate bars in the previous week (on average)

The median is *the value lying in the "middle" of the data set.*

- It separates the higher half from the lower half of a data sample
- It is calculated as follows:
    - Arrange the numbers in numerical order
    - Count how many numbers there are
    - If it is an odd number, divide by 2 and round up to get the position of the median number
    - If you have an even number, divide by 2. Go to the number in that position and average it with the number in the next higher position

Example 1: Dataset with 11 values: 0, 1, 2, 2, 3, **4**, 5, 6, 7, 9, 16

Median = 4 (4 is in the middle of the dataset), Mean = 5

=> one half of the people in the sample consumed 4 or less chocolate bars and the other half consumed 4 or more bars

Example 2: Dataset with 10 values: 1, 2, 2, 3, **4, 5**, 6, 7, 9, 16

Median is 4.5 (the average of 4 and 5 in the middle of the dataset), Mean = 5.5

=> In that case, half of the people in the sample ate less than 4.5 chocolate bars and the other half ate more than 4.5 bars

Example 3: Dataset with 11 values: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 93

Median = 4, Mean = 12

The median is not affected by extreme values or outliers (i.e., data points that differ significantly from other observations)

⇨ **The median is a "robust" measure of central tendency**

The mean is susceptible to outliers

⇨ **The mean is a "non-robust" measure of central tendency**

The median is often used in daily life, to represent the middle of a group.

- For example, when discussing the "average" income for a country, analyst will often use the median (rather than the mean) salary.
- As a robust measure of central tendency, the median is resistant to outliers (e.g., billionaires). It constitutes a fairer way to represent a "typical" income level.

The mode is the most commonly observed value in the data set. It is calculated as follows:

- Arrange the numbers in numerical order

- Count how many there are of each number.
- The number that appears most often is the mode.

Dataset with 11 values: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16

Mode = 2.

- The value "2" is represented twice, while all the other values are only represented once
- This means that 2 is the most common number of chocolate bars consumed in the previous week

Dataset with 11 values: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 16

- The series has no mode, as all values appear just once

Dataset with 11 values: 0, 1, 2, 2, 4, 4, 5, 6, 7, 9, 16

- The modes are 2 and 4, as they are both represented twice while all the other values are still only represented once
- This shows that a data set can have two (or more) modes

Dataset with 11 values:: 0, 1, 2, 2, 4, 4, 5, 6, 7, 9, 100

- The modes are 2 and 4 =>

The mode is a "robust" measure of central tendency.

The mode is not frequently used in daily or business life. However, it presents the key advantage that it can be used for nominal data (which is not possible with the mean and the median)

This means that it would be possible to calculate the mode if the data set showed the brand names of the chocolate bars purchased). Yet it makes no sense to speak of the "mean" or "median" brand

Measures of central tendency are particularly convenient when used to compare different variables or different subgroups. For example:

- To get a snapshot about the average (mean or median) income, wealth and tax rate of a population
- To compare the income, wealth and tax rate for different subgroups of a population (e.g., by region, gender, age group, or education)

N.B. A comparison of the *mean* and *median* can offer additional clues about the data

- A similar mean and median indicate that the values are quite balanced around the mean
- If, however, the median is lower than the mean, it is possible that there are outliers at the high end of the value spectrum (or "distribution" in the statistics jargon)

If median income < mean income => there is a large low- or middle-class population with a small minority with extremely high incomes (billionaires).

If median income > mean income => the economy probably consists of a large middle class and a small, extremely poor, minority.

4.4. Measures of spread

Measures of spread (also known as “measures of variability” and “measures of dispersion”) describe the dispersion of data within the data set. The dispersion is the extent to which data points depart from the center and from each other

The higher the dispersion, the more “scattered” the data points are.

Main measures of spread:

- the minimum and maximum,
- the range,
- the variance and standard deviation

Minimum and maximum - the lowest, respectively the highest values of the data set

Dataset with 11 values: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16

- The minimum = 0; the maximum = 16

A minimum or maximum that appears too low, respectively too high may suggest problems with the data set. The records related to these values should be carefully examined.

Range - the difference between the maximum and the minimum

- It is easy to calculate
- It provides a quick estimate about the spread of values in the data set, but is not a very good measure of variability
- In the previous example, the range is 16 (= 16-0)

Variance - describes how far each value in the data set is from the mean (and hence from every other value in the set)

- Mathematically, it is defined as the average of the squares of the differences between the observed and the mean
- It is always positive
- Due to its squared nature, the variance is not widely used in practice

Dataset: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16 - the variance is 18.73

Standard deviation - measures the dispersion of a data set relative to its mean

- It is calculated as the square root if the variance
- It expresses by how much the values of data set differ from the mean value for that data set
- In simple terms, it can be regarded as the average distance from the mean.
- A low standard deviation reveals that the values tend to be close to the mean of the data set
- Inversely, a high standard deviation signifies that the values are spread out over a wider range
- A useful property of the standard deviation compared to the variance is that it is expressed in the same unit as the data

Dataset: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16 - the standard deviation is 4.32

5. Interpreting Data

Different types of analyses or method produce various forms of output. These can be statistics, coefficients, probabilities, errors, etc., which can provide different insights to the reader.

It is important to be able to interpret these results, to understand what they mean, to know what kind of conclusions can be drawn and how to apply them for further decision making.

Data interpretation requires domain expertise, but also curiosity to ask the right questions. When the results are not in line with one's expectations, these should be met with a sufficient level of doubt and examined in further depth. Such sense of mistrust and inquisitiveness can be trained.

Five types of data interpretation approaches:

1) Correlation

2) Linear regression

3) Forecasting

4) Statistical tests

5) Classification

5.1. Correlation analysis

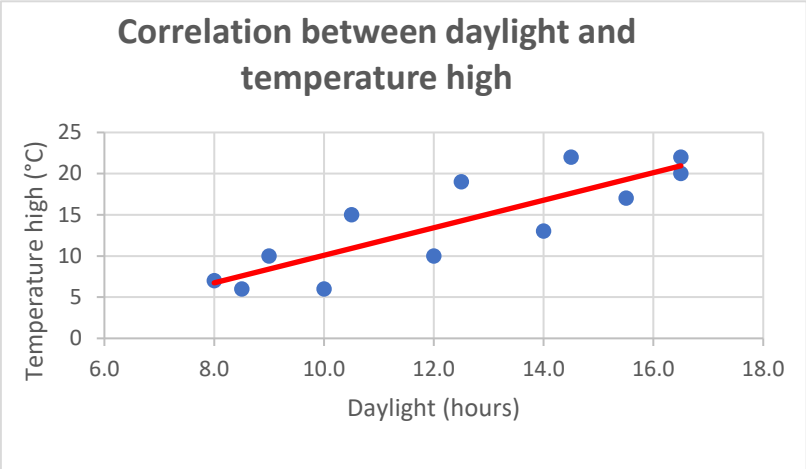
Correlation analysis is a *technique aimed at establishing whether two quantitative variables are related*

- It provides an indication of the direction and strength of the linear relationship between the two variables

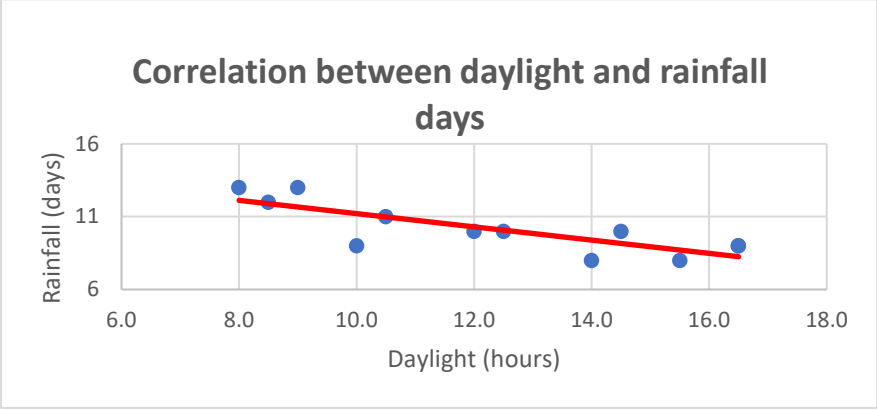
- It is only applicable with two numerical (quantitative) variables.
- Correlations between two variables can be presented graphically in a scatter plot:

When a **correlation** exists, you should be able to draw a straight line (called "regression line") that fits te data well.

Positive correlation (upward sloping regression line) - both variables move in the same direction. When one increases/decreases, the other increases/decreases as well.

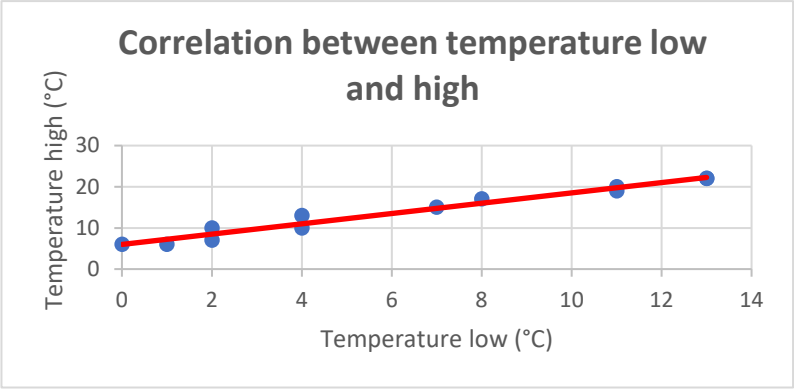


Negative correlation (downward sloping regression line) - the variables move in opposite directions. When one increases/decreases, the other decreases/increases.

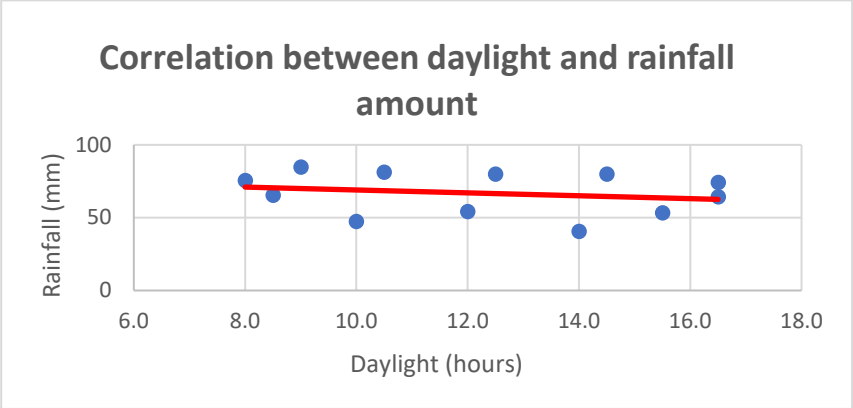


The more tightly the plot forms a line rising from left to right, the stronger the correlation.

"Perfect" correlation - all the points would lie on the regression line itself



Lack of correlation - If a line does not fit, i.e. if the dots are located far away from the line, it means that the variables are not correlated. In that case, the line is relatively flat.



5.1.1. Correlation coefficient

Although the visual examination of a scatter plot can provide some initial clues about the correlation between two variables, relying solely on them may lead to an interpretation that is too subjective.

It is also possible to determine the correlation of variables numerically, thus offering more precise evidence.

Correlation coefficient - a statistic that summarizes in a single number the relationship that can be seen in a scatter plot.

Basics about correlation coefficients:

The possible values of a correlation coefficient range from -1 to +1

Correlation coefficient = -1 => perfect negative correlation

Correlation coefficient = +1 => perfect negative correlation

When the value of one variable increases/decreases, the value of the other one moves in perfect sync.

The higher the absolute value of the correlation coefficient, the stronger the correlation

Strong positive (respectively negative) correlation, when the value of one variable increases/decreases, the value of the other variable increases (respectively decreases) in a similar manner

Correlation coefficient = 0 => there is no linear relationship between the two variables.

However, this does not necessarily mean the variables are not related at all. They may have some other form of relationship (e.g., an exponential or logarithmic relationship), but not a linear one.

Correlation in the real world rarely return coefficients of exactly +1.0, –1.0, or 0. Most of the time, they fall somewhere in between.

A rule of thumb for the interpretation of the absolute value of correlation coefficients:

- Below 0.25: No correlation
- Between 0.25 and 0.50: Weak correlation
- Between 0.50 and 0.75: Moderate correlation
- Above 0.75: Strong correlation

When applying this rule of thumb, always bear in mind that the definition of a weak, moderate, or strong correlation may vary from one domain to the next.

5.1.2. Correlation and causation

Correlation ≠ causal relationship.

Correlation only establishes a statistical relationship. No matter how strong or significant the correlation is, however, it never provides sufficient evidence to claim a causal link between variables. This applies to both the *existence* or the *direction* of a cause-and-effect relationship, for which no conclusion can made whatsoever.

The strong correlation between two variables A and B can be due to various scenarios:

- Direct causality: A causes B
- Reverse causality: B causes A
- Bidirectional causality: A causes B and B causes A
- A and B are both caused by C
- Pure coincidence, i.e., there is no connection between A and B

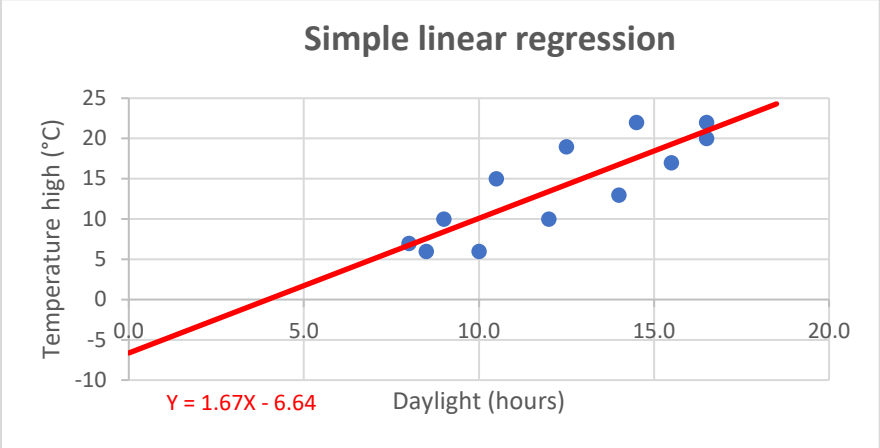
Determining an actual cause-and-effect relationship between the variables requires further investigation.

The presence of correlation, combined with sound business judgment, is sometimes enough to make quality decisions. A good analyst or a wise decision maker will prefer to strive for an explanation and always bear in mind the old statistical adage:

Correlation does not imply causation.

5.2. Simple linear regression

With simple linear regression, it is possible to predict the value of the **dependent variable** based on the value of one single independent variable.



Regression equation: Y = a + bX

Y (dependent variable) - the one to be explained or to be predicted

X (independent variable, predictor) – the one explaining or predicting the value of Y

a (intercept) - a constant, corresponding to the point at which the line crosses the vertical axis, i.e., when X is equal to zero

b (slope) - the coefficient of X, quantifying how much Y changes for each incremental (one-unit) change in X.

N.B. The higher the absolute value of b, the steeper the regression curve

The sign of b indicates the direction of the relationship between the Y and X

b > 0 - the regression line shows an upward slope. An increase of X results in an increase of Y

b < 0 - the regression line shows a downward slope. An increase of X results in a decrease of Y

Notice that a prediction using simple linear regression does not prove any causality. The coefficient b, no matter its absolute value, says nothing about a causal relationship between the dependent and the independent variables.

Summary: The objective of the regression is to plot the one line that best characterizes the cloud of dots.

5.2.1. R-squared

After running a linear regression analysis, you need to examine how well the model fits the data., i.e., determine if the regression equation does a good job explaining changes in the dependent variable.

*The regression model fits the data well when the differences between the observations and the predicted values are relatively small. If these differences are too large, or if the model is biased, you cannot trust the results***.**

R-squared (R^2, coefficient of determination) - estimates the scatter of the data points around the fitted regression line

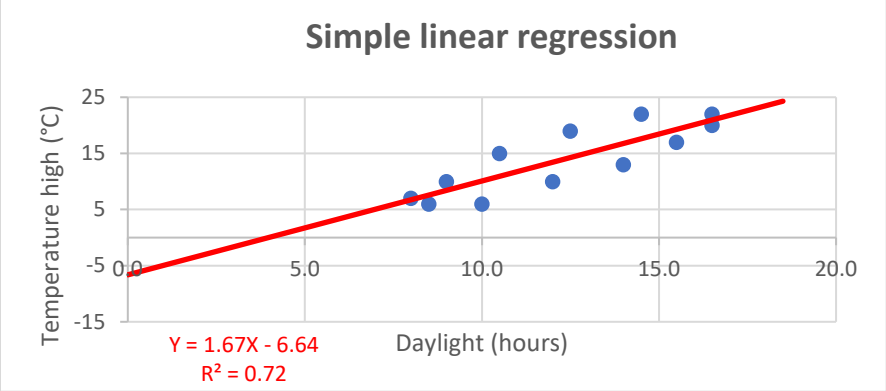
It represents the proportion of the variance in a dependent variable that can be explained by the independent variable.

The remaining variance can be attributed to additional, unknown, variables or inherent variability.

In simple linear regression models, the R-squared statistic is always a number between 0 and 1, respectively between 0% and 100%

- **R-squared = 0** => the model does not explain any of the variance in the dependent variable. The model is useless and should not be used to make predictions
- **R-squared = 1** => the model explains the whole variance in the dependent variable. Its predictions perfectly fit the data, as all the observations fall exactly on the regression line.

Example:



R-squared = 0.72

=> the number of daylight accounts for about 72% of the variance of the temperature

=> the number of daylight hours is a good predictor of the temperature.

What is a good R-squared value? At what level can you trust a model?

- Some people or textbooks claim that 0.70 is such a threshold, i.e., that if a model returns an R-squared of 0.70, it fits well enough to make predictions based on it. Inversely, a value of R-squared below 0.70 indicates that the model does not fit well.
- Feel free to use this rule of thumb. At the same time, you should be aware of its caveats.
- Indeed, the properties of R-squared are not as clear-cut as one may think

A higher R-squared indicates a better fit for the model, producing more accurate predictions

**A model that explains 70% of the variance is likely to be much better than one that explains 30% of the variance. However, such a conclusion is not necessarily correct.*

R-squared depends on various factors, such as:

- The sample size: The larger the number of observations, lower the R-squared typically gets
- The granularity of the data: Models based on case-level data have lower R-squared statistics than those based on aggregated data (e.g., city or country data)

- The type of data employed in the model: When the variables are categorical or counts, the R-squared will typically be lower than with continuous data
- The field of research: studies that aim at explaining human behavior tend to have lower R-squared values than those dealing with natural phenomena. This is simply because people are harder to predict than stars, molecules, cells, viruses, etc.

5.3. Forecasting

Forecast accuracy – the closeness of the forecasted value to the actual value.

For a business decision maker, the key question is how to determine the accuracy of forecasts:

"Can you trust the forecast enough to make a decision based on it?

As the actual value cannot be measured at the time the forecast, the accuracy can only be determined retrospectively.

Accuracy of forecasts is tested by calculating different statistical indicators, also known as errors. These errors can be assessed without knowing anything about a forecast except its past values.

The three most common forecast errors:

1) the Mean Absolute Error (MAE)

2) the Mean Absolute Percent Error (MAPE)

3) the Root Mean Square Error (RMSE)

**The MAE, MAPE and RMSE only measure typical errors. They cannot anticipate black swan events like financial crises, global pandemics, terrorist attacks, or the Brexit.*

As the errors associated with these events are not covered by the time series data, they cannot be modeled. Accordingly, it is impossible to determine in advance how big the error will be.

5.3.1. Forecast errors

Mean Absolute Error (MAE) - the absolute value of the difference between the forecasted values and the actual value.

- With the MAE, we can get an idea about how large the error from the forecast is expected to be on average.
- The main problem with it is that it can be difficult to anticipate the relative size of the error. How can we tell a big error from a small error?

What does a MAE of 100 mean? Is it a good or bad forecast?

That depends on the underlying quantities and their units. For example, if your monthly average sales volume is 10'000 units and the MAE of your forecast for the next month is 100, then this is an amazing forecasting accuracy. However, if sales volume is only 10 units on average, then we are talking of a rather poor accuracy.

Mean Absolute Percentage Error (MAPE) - the sum of the individual absolute errors divided by the underlying value.

- It corresponds to the average of the percentage errors
- The MAPE allows us to compare forecasts of different series in different scales.

- Hence, we could, for example, compare the accuracy of a forecast of sales value and volume in US$, EUR, units, liters or gallons - even though these numbers are in different units.
- Thanks to this property, the MAPE is one of the most used indicators to measure forecast accuracy
- One key problem with the MAPE is that it may understate the influence of big, but rare, errors. Consequently, the error is smaller than it should be, which could mislead the decision maker.

Root Mean Square Error (RMSE) - the square root of the average squared error.

RMSE has the key advantage of giving more importance to the most significant errors. Accordingly, one big error is enough to lead to a higher RMSE. The decision maker is not as easily pointed in the wrong direction as with the MAE or MAPE.

- Best practice is to compare the MAE and RMSE to determine whether the forecast contains large errors. The smaller the difference between RMSE and MAE, the more consistent the error size, and the more reliable the value.
- As with all "errors", the objective is always to avoid them.

The smaller the MAE, MAPE or RMSE, the better!

What is a "good" or "acceptable" value of the MAE, MAPE or RSME depends on:

- The environment: In stable environment, in which demand or prices do not vary so much over time (e.g., electricity or water distribution), demand or sales volumes are likely to be rather steady and predictable. A

forecasting model may therefore yield a very low MAPE, possibly under 5%.

- The industry: In volatile industries (e.g., machine building, oil & gas, chemicals) or if the company is exposed to hypercompetition and constantly has to run advertising campaigns or price promotions (like in the FMCG or travel industries), sales volumes vary significantly over time and are much more difficult to be forecasted accurately. Accordingly, the MAPE of a model could be much higher than 5%, and yet be useful for decision makers in the sales, finance or supply chain departments.
- The type of company: Forecasts for larger geographic areas (e.g., continental or national level) are generally more accurate than for smaller areas (e.g., regional or local).
- The time frame: Longer period (e.g., monthly) forecasts usually yield higher accuracies than shorter period (e.g., daily or hourly) forecasts.

With three well-established indicators available, one cannot conclude that one is better than the other. Each indicator can help you avoid some shortcomings but will be prone to others. Only experimentation with all three indicators can tell you which one is best, depending on the phenomenon to be forecasted.

5.4. Statistical tests

Hypothesis testing is a key tool in inferential statistics, and used in various domains - social sciences, medicine, and market research. The purpose of hypothesis

testing is to establish whether there is enough statistical evidence in favor of a certain idea or assumption, i.e., the hypothesis.

The process involves testing an assumption regarding a population by measuring and analyzing a random sample taken from that population

Population - the entire group that is being examined.

If you want to compare the satisfaction levels of male and female employees in your company, the population is made of the entire workforce (25,421 people)

Sample - the specific group that data are collected from. Its size is always smaller than that of the population.

If you randomly select 189 men and 193 women among these employees to carry out a survey, these 382 employees constitute your sample.

Hypothesis testing becomes particularly relevant when census data cannot be collected, for example because the process would be too lengthy or too expensive. In these cases, researchers need to develop specific experiment designs, and rely on survey samples to collect the necessary data.

Modern statistical software is there to calculate various relevant statistics, test values, probabilities, etc. All you need to do is to learn interpret the most important ones:

- null hypothesis
- the p-value
- statistical significance.

5.4.1. Hypothesis testing

A hypothesis resembles a theory in science. But it is “less” than that, because it first needs to go through extensive testing before it can be deemed a proper theory.

Hypotheses are formulated as statements, not as questions.

A hypothesis constitutes a starting point for further analytical investigation.

Step 1 in any statistical test is to define a hypothesis, which is a statement that helps communicate an understanding of the question or issue at stake. It is common to propose two opposite, mutually exclusive, hypotheses so that only one can be right: *The null hypothesis (Ho) and the alternative hypothesis (H1).*

The null hypothesis (Ho)

The null hypothesis represents the commonly accepted fact

It is usually expressed as a hypothesis of "no difference" in a set of given observations, for example:

- *A blue conversion button on the website results in the same CTR as a red button*

In statistics terms, the null hypothesis is therefore usually stated as the equality between population parameters. For example:

- *The mean CTRs of the red and blue conversion buttons are the same.*

OR: the difference of the mean CTRs of the red and the blue conversion buttons is equal to zero

It is called "null" hypothesis, because it is usually the hypothesis that we want to nullify or to disprove.

The alternative hypothesis (H1) is the one that you want to investigate, because you think that it can help explain a phenomenon

It represents what you believe to be true or hope to prove true.

- In the example above, the alternative hypothesis could be formulated as follows:
  - *A blue conversion button on the website will lead to a different CTR than the one with a red button*

In this case, the objective is to determine whether the population parameter is generally distinct or differs in either direction from the hypothesized value. It is called a two-sided (or non-directional) alternative hypothesis.

Sometimes, it can be useful to determine whether the population parameter differs from the hypothesized value in a specific direction, i.e. is smaller or greater than the value. This is known as a one-sided (or directional) alternative hypothesis

- *Example: The difference of the mean CTRs of the blue and the red conversion button is positive.*
- *Here we only care about the blue button yielding a higher CTR than the red button*

We can also be even more aggressive in our statement and quantify that difference, for example:

- The difference of the mean CTRs of the red and the blue conversion button is higher than 2 percentage points

- That would be equivalent to stating that the mean CTR of the blue button is 2 percentage points higher than mean CTR of the red button

N.B. You do not have to specify the alternative hypothesis. Given that the two hypotheses are opposites and mutually exclusive, only one can, and will, be true. For the purpose of statistical testing, it is enough to reject the null hypothesis. It is therefore very important to work out a clear null hypothesis.

5.4.2. P-value

P-value is a measure of the probability that an observed result could have occurred just by random chance, as opposed to a certain pattern.

The greater the dissimilarity between these patterns, the less likely it is that the difference occurred by chance.

Examples:

If p-value = 0.0326 => there is a 0.0326 (or 3.26%) chance that the results happened randomly.

If p-value = 0.9429 => the results have a 94.29% chance of being random

The smaller the p-value, the stronger the evidence that you should reject the null hypothesis

When you see a report with the results of statistical tests, look out for the p-value. Normally, the closer to 0.000, the better – depending, of course, on the hypotheses stated in that report.

5.4.3. Statistical significance

The **significance level** (alpha, α) is a number stated in advance to determine how small the p-value must be to reject the null hypothesis.

- The researcher sets it arbitrarily, before running the statistical test or experiment
- If the p-value falls below the significance level, the result of the test is statistically significant.
- Unlike the p-value, the alpha does not depend on the underlying hypotheses, nor is it derived from any observational data.
- The alpha will often depend on the scientific domain the research is being carried out in

If p-value < alpha => you can reject the null hypothesis at the level alpha.

If the P-value is lower than the significance level alpha, which should be set in advance, then we can conclude that the results are strong enough to reject the old notion (the null hypothesis) in favor of a new one (the alternative hypothesis).

Example:

If p-value = 0.0321, alpha = 0.05 => "Based on the results, we can reject the null hypothesis at the level of significance $\alpha = 0.05$ (as $p = 0.0321 < 0.05$).

If the p-value = 0.1474, alpha = 0.05 => "Based on the results, we accept the null hypothesis at the level of significance $\alpha = 0.05$ (as $p = 0.1474 >= 0.05$).

N.B. Statistical significance ≠ practical (theoretical) significance.

A result that is statistically significant is not necessarily "meaningful" or "important". That will depend on the real-world relevance of that result, which the researcher must determine.

5.5. Classification

A classification model can only achieve two results: Either the prediction is correct (i.e., the observation was placed in the right category), or it is incorrect.

This characteristic makes it rather straightforward to estimate the quality of a classification model, especially when there are only two available categories or labels.

Out-of-sample validation - withholding some of the sample data used for the training of the model. Once the model is ready, it is validated with the data initially set aside for this very purpose.

Example:

Imagine that we trained a model for a direct marketing campaign. We used the data available to predict each recipient's response to the marketing offer:

- "yes"- favorable response
- "no" – negative response

We set aside 100 customer records, which constitute our validation data.

- For these 100 customers, we use the model to predict their responses. These constitute the predicted classes.

- As these customers also receive the marketing offer, we also get to know who responded favorably, and who did not. These responses constitute the actual classes.
- You can compare the predicted with the actual classes, and find out which predictions were correct.

Confusion matrix - shows the actual and predicted classes of a classification problem (correct and incorrect matches). The rows represent the occurrences in the actual class, while the columns represent the occurrences in the predicted class.

n = 100		**Predicted class**		
		Yes	**No**	
Actual class	**Yes**	10	5	**15**
	No	15	70	**85**
		25	**75**	**100**

There are two possible responses to a marketing offer:

- "yes" - the customers accept it
- "no" - they ignore or reject it.

Out of the 100 customer who received the offer, the model predicted that 25 customers would accept it (i.e., 25 times “yes”) and that 75 customers would reject it (i.e., 75 times “No”

After running the campaign, it turned out that 15 customers responded favorably (“Yes”), while 85 customers ignored it (“No”).

Based on the confusion matrix, one can estimate the quality of a classification model by calculating its:

- **Accuracy**
- **Recall**
- **Precision**

5.5.1. Accuracy

Accuracy is the proportion of the total number of correct predictions.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ total\ observations}$$

The model correctly predicted 10 "Yes" cases and 70 "No" cases =>

$$Accuracy = \frac{10 + 70}{100} = 80\%$$

- Accuracy is the most fundamental metric used to assess the performance of classification models.
- It is intuitive and widely used.
- It comes with a major flaw, which becomes apparent when the classes are imbalanced.
- Experienced analysts are familiar with this issue, and have at their disposal various techniques to handle imbalanced data sets

5.5.2. Recall and precision

Recall (also known as sensitivity) is the ability of a classification model to identify *all* relevant instances.

$$Recall = \frac{Total\ correct\ predictions}{Actual\ positive\ cases}$$

The model correctly predicted 10 out of 15 positive responses =>

$$Recall = \frac{10}{15} = 66.67\%$$

- This means that only two-thirds of the positives were identified as positive, which is not a very good score.

Precision is the ability of a classification model to return *only* relevant instances (to be correct when predicting "yes").

$Precision = \frac{Total\ correct\ predictions}{Predicted\ positive\ cases}$The model predicted 25 positive responses, out of which 10 were correctly predicted =>

The model predicted 25 positive responses, out of which 10 were correctly predicted =>

$$Precision = \frac{10}{25} = 40\%$$

There are two types of incorrect predictions: false positives and false negative

A false positive - when a case is labeled as positive although it is actually negative

- The model predicts "yes" where the case is actually "no"
- When the objective is to minimize false positives, it is advisable to assess classification models using precision.

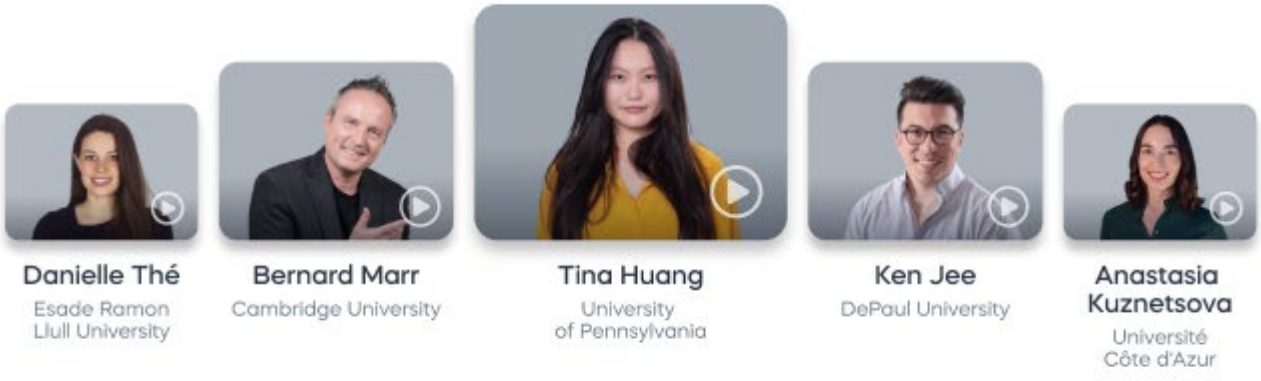
A false negative - when a case is labeled as negative although it is actually positive

- The model predicts "no" where the case is actually "yes"
- When the objective is to minimize false negatives, it is advisable to assess classification models using recall.

Learn **DATA SCIENCE** anytime, anywhere, at your own pace.

If you found this resource useful, check out our **e-learning program**. We have everything you need to succeed in data science.

Learn the most sought-after data science skills from the **best experts in the field**! Earn a **verifiable certificate** of achievement trusted by employers worldwide and future proof your car



Danielle Thé
Esade Ramon Llull University

Bernard Marr
Cambridge University

Tina Huang
University of Pennsylvania

Ken Jee
DePaul University

Anastasia Kuznetsova
Université Côte d'Azur

Comprehensive training, exams, certificates.

- ✓ 160+ hours of video
- ✓ 599+ Exercises
- ✓ Downloadables
- ✓ Exams & Certification
- ✓ Personalized support
- ✓ Resume Builder & Feedback
- ✓ Portfolio advice
- ✓ New content
- ✓ Career tracks

Join a global community of 1.8 M successful students with an annual subscription at 60% OFF with coupon code **365RESOURCES**.

~~$432~~ **$172.80**/year

Start at 60% Off

VAT may be applied

365 DataScience

Olivier Maugain

Email: team@365datascience.com