UNIT-1
INTRODUCTION

SYLLABUS: UNIT I

Introduction: Need for data science – benefits and uses – facets of data – data science process – setting their search goal – retrieving data – cleansing, integrating, and transforming data – exploratory data analysis – build the models – presenting and building applications.

Data Science

→ *Data science* involves using methods to analyze massive amounts of data and extract the knowledge it contains.

→ Data science is an evolutionary extension of statistics capable of dealing with the massive amounts of data produced today. It adds methods from computer science to the repertoire of statistics.

→ Data Science is a discipline that merges concepts from computer science (algorithms, programming, machine learning, and data mining), mathematics (statistics and optimization), and domain knowledge (business, applications, and visualization) to extract insights from data and transform it into actions that have an impact in the particular domain of application.

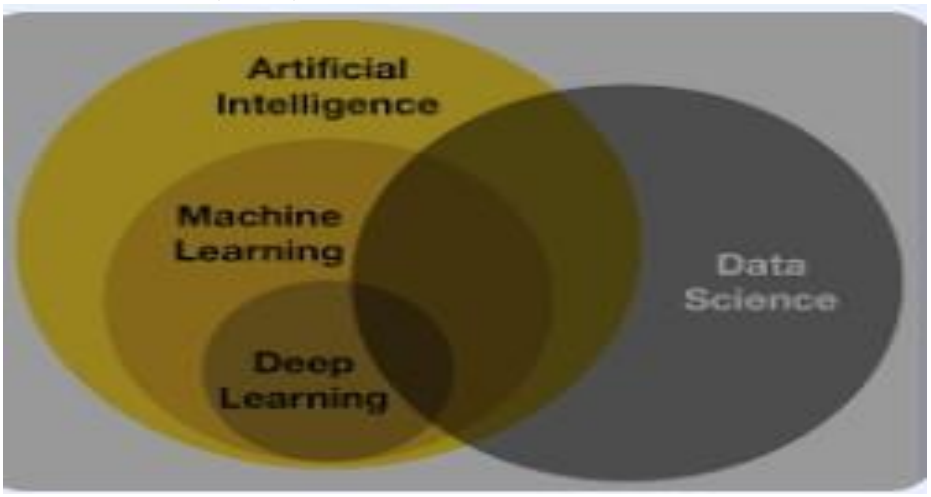
→ Data science is the field of study that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems that perform tasks which ordinarily require human intelligence. In turn, these systems generate insights that analysts and business users translate into tangible business value.

→ Data science is a field that studies data and how to extract meaning from it, using a series of methods, algorithms, systems, and tools to extract insights from structured and unstructured data. That knowledge then gets applied to business, government, and other bodies to help drive profits, innovate products and services, build better infrastructure and public systems, and more.

→ *In short, Data Science* "uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms".

→ Python is a great language for data science because it has many data science libraries available, and it's widely supported by specialized software. For instance, almost every popular NoSQL database has a Python-specific API. Because of these features and the ability to prototype quickly with Python while keeping acceptable performance, its influence is steadily growing in the data science world.

→ Relation between AI, ML, DL and data science



Need for Data Science

→ The principal purpose of Data Science is to find patterns within data. It uses **various statistical techniques** to analyze and draw insights from thedata.

→ Every business has data but its business value depends on how much they know about the data they have.

→ Data Science has gained importance in recent times because it can help businesses to increase business value of its available data which in turn can help them to take competitive advantage against their competitors.

→ It can help us to know our customers better, it can help us to optimize our processes, it can help us to take better decisions. Because of data science, data has become strategic asset.

→ *Data creates magic*. Industries need data to help them make careful decisions. Data Science churns raw data into meaningful insights. Therefore, industries need data science.

→ A Data Scientist is a wizard who knows how to create magic using data.

→ By using Data Science, companies are able to make:

- Better decisions (should we choose A or B)
- Predictive analysis (what will happen next?)
- Pattern discoveries (find pattern, or maybe hidden information in the data)

Benefits and Uses of Data Science

→ Data science is used almost everywhere in both commercial and non commercial settings.

→ Commercial companies in almost every industry use data science to gain insights into their customers, processes, staff, completion, and products.

→ Many companies use data science to offer customers a better user experience, as well as to cross-sell, up-sell, and personalize their offerings.

→ A good example of this is Google AdSense, which collects data from internet users so relevant commercial messages can be matched to the person browsing the internet. MaxPoint is another example of real-time personalized advertising.

→ Human resource professionals use people analytics and text mining to screen candidates, monitor the mood of employees, and study informal networks among co workers.

→ Financial institutions use data science to predict stock markets, determine the risk of lending money, and learn how to attract new clients for their services.

→ Governmental organizations are also aware of data's value. Many governmental organizations not only rely on internal data scientists to discover valuable information, but also share their data with the public. We can use this data to gain insights or build data-driven applications. *Data.gov* is but one example; it's the home of the US Government's open data.

→ Universities use data science in their research but also to enhance the study experience of their students. The rise of massive open online courses (MOOC) produces a lot of data, which allows universities to study how this type of learning can complement traditional classes. MOOCs are an invaluable asset if we want to become a data scientist and big data professional, so definitely look at a few of the better-known ones: Coursera, Udacity, and edX.

Facets of data

(Types of Data or Different Forms of Data)

→ Data can come in different forms. The main forms are

- Structured data
- Unstructured data
- Natural language data
- Machine-generated data
- Graph-based data or Network data
- Audio, video, and images
- Streaming data

Structured Data:

→ Structured data is data that depends on a data model and resides in a fixed field within a record.

→ As such, it's often easy to store structured data in tables within databases or Excel files.

→ SQL, or Structured Query Language, is the preferred way to manage and query data that resides in databases.
→ More often, data comes unstructured.

Unstructured Data

→ Unstructured data is data that isn't easy to fit into a data model because the content is context-specific or varying.
→ One example of unstructured data is our regular email.
→ Although email contains structured elements such as the sender, title, and body text, it's a challenge to find the number of people who have written an email complaint about a specific employee because so many ways exist to refer to a person, for example. The thousands of different languages and dialects out there further complicate this.
→ A human-written email is also a perfect example of natural language data.

Natural language Data:

→ Natural language is a special type of unstructured data; it's challenging to process because it requires knowledge of specific data science techniques and linguistics.
→ The natural language processing community has had success in entity recognition, topic recognition, summarization, text completion, and sentiment analysis, but models trained in one domain don't generalize well to other domains.
→ Even state-of-the-art techniques aren't able to decipher the meaning of every piece of text. This shouldn't be a surprise though: humans struggle with natural language as well. It's ambiguous by nature.
→ The concept of meaning itself is questionable here. Have two people listen to the same conversation. Will they get the same meaning? The meaning of the same words can vary when coming from someone upsetor jowes.

Machine-generated Data:

→ Machine-generated data is information that's automatically created by a computer, process, application, or other machine without human intervention.
→ Machine-generated data is becoming a major data resource and will continue to do so.
→ Wikibon has forecast that the market value of the *industrial Internet* (a term coined by Frost & Sullivan to refer to the integration of complex physical machinery with networked sensors and software) will be approximately $540 billion in 2020.
→ IDC (International Data Corporation) has estimated there will be 26 times more connected things than people in 2020. This network is commonly referred to as *the internet of things*.

→ The analysis of machine data relies on highly scalable tools, due to its high volume and speed.

→ Examples of machine data are web server logs, call detail records, network event logs, and telemetry.

Graph-based or Network Data:

→ "Graph data" can be a confusing term because any data can be shown in a graph. "Graph" in this case points to mathematical graph theory.

→ In graph theory, a graph is a mathematical structure to model pair-wise relationships between objects. Graph or network data is, in short, data that focuses on the relationship or adjacency of objects.

→ The graph structures use nodes, edges, and properties to represent and store graphical data.

→ Graph-based data is a natural way to represent social networks, and its structure allows calculating specific metrics such as the influence of a person and the shortest path between two people.

→ Examples of graph-based data can be found on many social media websites. For instance, on LinkedIn we can see who we know at which company. Our follower list on Twitter is another example of graph-based data.

→ The power and sophistication comes from multiple, overlapping graphs of the same nodes. For example, imagine the connecting edges here to show "friends" on Face book. Imagine another graph with the same people which connect business colleagues via LinkedIn. Imagine a third graph based on movie interests on Netflix. Overlapping the three different- looking graphs makes more interesting questions possible.

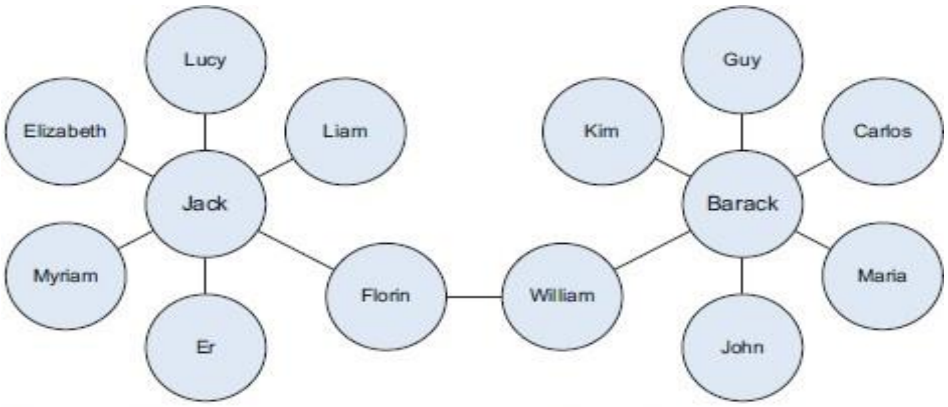


Figure 1.4 Friends in a social network are an example of graph-based data.

Audio, image, and video Data:

→ Audio, image, and video are data types that pose specific challenges to a data scientist.

→ Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.

→ MLBAM (Major League Baseball Advanced Media) announced in 2014 that they'll increase video capture to approximately 7 TB per game for the purpose of live, in-game analytics. High-speed cameras at stadiums will capture ball and athlete movements to calculate in real time, for example, the path taken by a defender relative to two baselines.

→ Recently a company called DeepMind succeeded at creating an algorithm that's capable of learning how to play video games. This algorithm takes the video screen as input and learns to interpret everything via a complex process of deep learning.

→ It's a remarkable feat that prompted Google to buy the company for their own Artificial Intelligence (AI) development plans. The learningalgorithm takes in data as it's produced by the computer game; it's streaming data.

Streaming Data

→ While streaming data can take almost any of the previous forms, it has an extra property. The data flows into the system when an event happens instead of being loaded into a data store in a batch. Although this isn't really a different type of data, we treat it here as such because we need to adapt process to deal with this type of information.

→ Examples are the "What's trending" on Twitter, live sporting or music events, and the stock market.

Data Science Process

(Steps in Data science process or Flow of Data science Process or phases of data science process)

→ The data science process typically consists of six steps.

1. Setting the research goal
2. Gathering data (or) Retrieving data
3. Data preparation (or) Data Pre processing
4. Data exploration (or) Exploratory data analysis
5. Data Modeling (or) Model building
6. Presentation and Automation

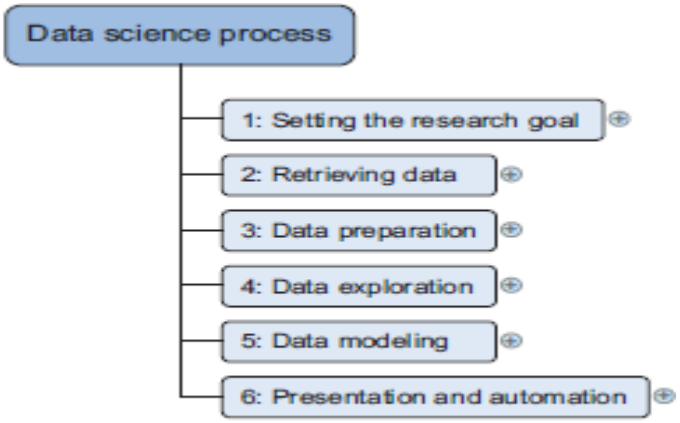


Figure 1.5 The data science process

Brief explanation of the six steps of the data science process

→ 1. The first step of this process is setting a *research goal*. The main purpose here is making sure all the stakeholders understand the *what*, *how*, and *why* of the project. In every serious project this will result in a project charter.

→ 2. The second phase is *data retrieval*. This step includes finding suitable data and getting access to the data owner. The result is data in its raw form, which probably needs polishing and transformation before it becomes usable.

→ 3. This includes transforming the data from a raw form into data that's directly usable in models. To achieve this, we'll detect and correct different kinds of errors in the data, combine data from different data sources, and transform it.

→ 4. The fourth step is *data exploration*. The goal of this step is to gain a deep understanding of the data. We'll look for patterns, correlations, and deviations based on visual and descriptive techniques.

→ 5. Finally, we get to the sexiest part: *model building* (often referred to as "data modeling"). It is now that we attempt to gain the insights or make the predictions stated in project charter.

→ 6. The last step of the data science model is *presenting results and automating the analysis*, if needed. Certain projects require performingthe business process over and over again, so automating the project will save time.

Advantages of this approach:

→ Following these six steps pays off in terms of a higher project success ratio and increased impact of research results. This process ensures we have a well-defined research plan, a good understanding of the business question, and clear deliverables before we even start looking at data.

→ Another benefit of following a structured approach is that we work more in *prototype mode* while we search for the best model.

→ Dividing a project into smaller stages also allows employees to work together as a team.

Six steps of the data science process

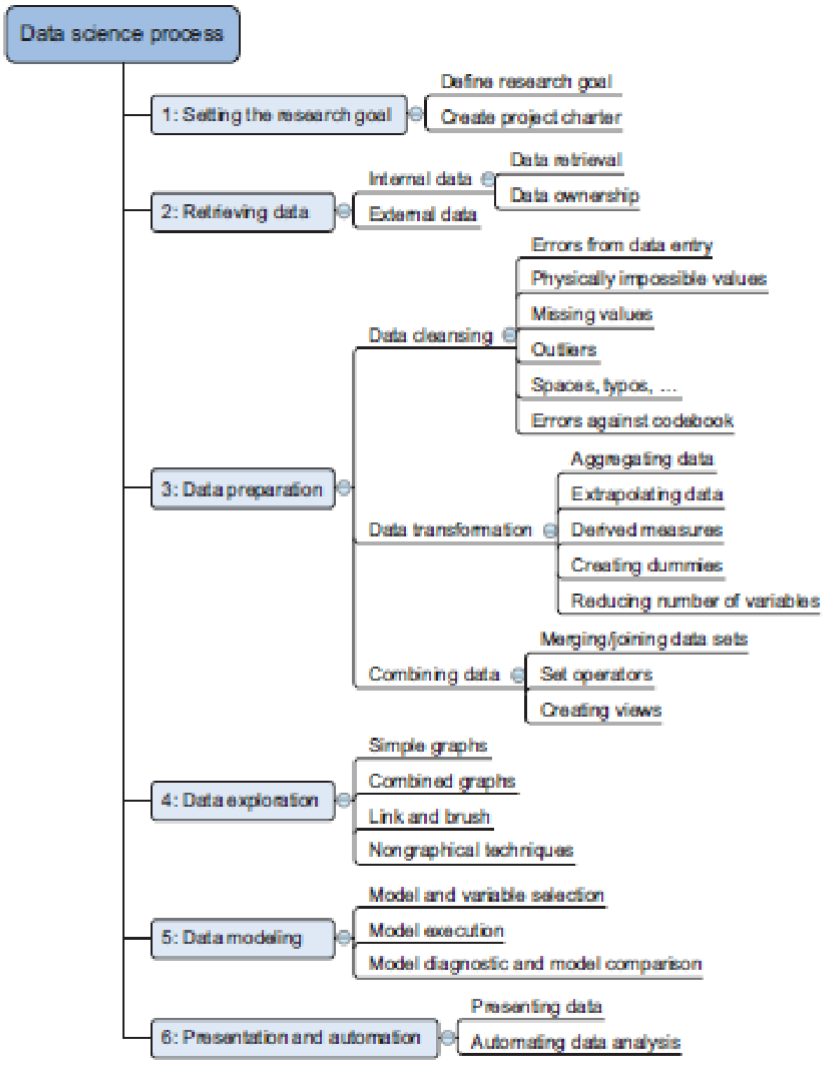


Figure 2.1 The six steps of the data science process

Step 1: Defining research goals and creating a project charter

→ A project starts by understanding the what, the why, and the how of project.

→ What does the company expect to do? And why does management place such a value on research? How the company benefits from that? Answering these three questions (what, why, how) is the goal of the first phase, so that everybody knows what to do and can agree on the best course of action.

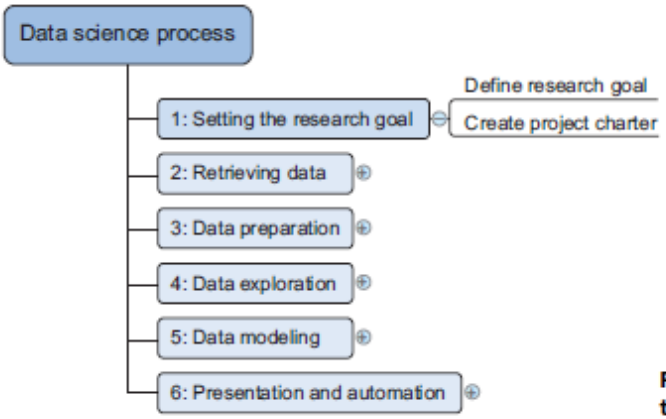


Figure 2.2 Step 1: Setting the research goal

→ The outcome should be a clear research goal, a good understanding of the context, well-defined deliverables, and a plan of action with a timetable. This information is then best placed in a project charter.

→ After we have a good understanding of the business problem, try to get a formal agreement on the deliverables. All this information is best collected in a project charter. For any significant project this would be mandatory.

→ A project charter requires teamwork, and the input covers at least the following:

- A clear research goal
- The project mission and context
- How we're going to perform analysis
- What resources we expect to use
- Proof that it's an achievable project, or proof of concepts
- Deliverables and a measure of success
- A timeline

Step 2: Retrieving data

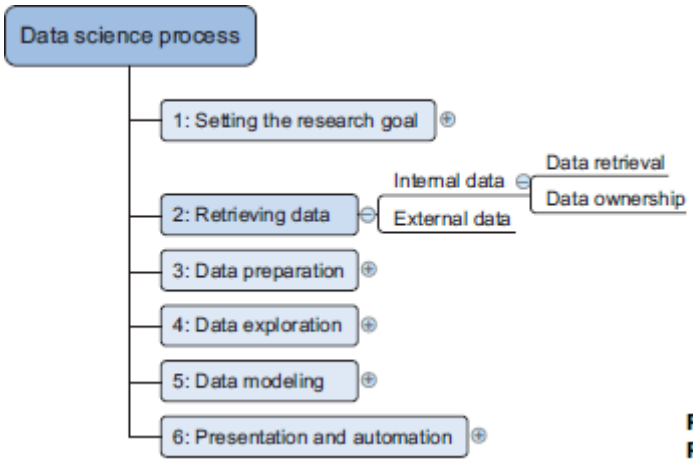


Figure 2.3 Step 2: Retrieving data

→ The next step in data science is to retrieve the required data. This data is either found within the company or retrieved from a third party.

→ The mail goal of this phase is finding suitable data and getting access to the data owner.

→ Data can be stored in many forms, ranging from simple text files to tables in a database. The data can be stored in official data repositories such as *databases*, *data marts*, *data warehouses*, and *data lakes* maintained by a team of IT professionals.

→ Finding data even within own company can sometimes be a challenge. As companies grow, their data becomes scattered around many places. Knowledge of the data may be dispersed as people change positions and leave the company

→ Getting access to data is another difficult task.. Getting access to the data may take time and involve company politics.

→ Although data is considered an asset more valuable than oil by certain companies, more and more governments and organizations share their data for free with the world. This data can be of excellent quality.

Step 3: Cleansing, integrating, and transforming data

→ This phase sanitize and prepare the data received from the data retrieval phase, for use in the modeling and reporting phase.

→ This phase involves, checking and remediating data errors, enriching the data with data from other data sources, and transforming it into a suitable format for models.

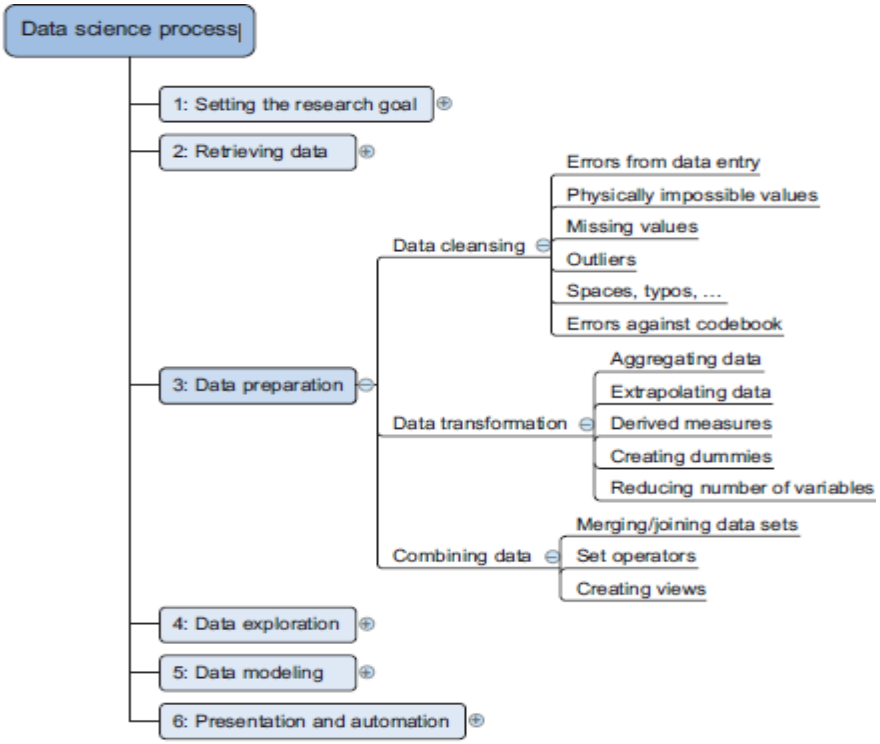


Figure 2.4 Step 3: Data preparation

→ This phase consist of the following main steps:

- Data cleaning
- Data transformation
- Data Integration (Combining data)

Cleansing data

→ Data cleansing is a subprocess of the data science process that focuses on removing errors in data so that data becomes a true and consistent representation of the processes it originates from.

→ By "true and consistent representation" we imply that at least two types of errors exist. The first type is the *interpretation error*, such "a person's age is greater than 300 years."

→ The second type of error points to *inconsistencies* between data sources . An example of this class of errors is putting "Female" in one table and "F" in another when they represent the same thing: that the person is female. Another example is that we use Pounds in one table and Dollars in another.

Table 2.2 An overview of common errors

General solution	
Try to fix the problem early in the data acquisition chain or else fix it in the program.	
Error description	**Possible solution**
Errors pointing to false values within one data set	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
Errors pointing to inconsistencies between data sets	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation

Data Entry Errors

→ Data collection and data entry are error-prone processes. They often require human intervention; they make typos or lose their concentration for a second and introduce an error into the chain.

→ But data collected by machines or computers isn't free from errors either. Errors can arise from human sloppiness, whereas others are due to machine or hardware failure. Examples of errors originating from machines are transmission errors or bugs in the extract, transform, and load phase (ETL).

→ For small data sets we can check every value by hand. Detecting data errors when the variables we study don't have many classes can be done by tabulating the data with counts.

Redundant Whitespace

→ Whitespaces tend to be hard to detect but cause errors like other redundant characters.

→ If we know to watch out for them, fixing redundant whitespaces is luckily easy enough in most programming languages. They all provide string functions that will remove the leading and trailing whitespaces. For instance, in Python we can use the **strip()** function to remove leading and trailing spaces.

Impossible Values and Sanity Checks

→ Sanity checks are another valuable type of data check.

→ Here we check the value against physically or theoretically impossible values such as people taller than 3 meters or someone with an age of 299 years.

→ Sanity checks can be directly expressed with rules: check = 0 <= age <= 120

Outliers

→ An outlier is an observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations.

→ The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.

→ Outliers can gravely influence the data modeling, so investigate them first.

Dealing with Missing Values

→ Missing values aren't necessarily wrong, but still need to handle them separately.

→ Certain modeling techniques can't handle missing values.

→ They might be an indicator that something went wrong in data collection or that an error happened in the ETL process.

→ Common techniques data scientists use are listed in following table:

Table 2.4 An overview of techniques to handle missing data

Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to `null`	Easy to perform	Not every modeling technique and/or implementation can handle `null` values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions

Deviations from a Code Book

→ Detecting errors in larger data sets against a code book or against standardized values can be done with the help of set operations.

→ A code book is a description of data, a form of metadata. It contains things such as the number of variables per observation, the number of observations, and what each encoding within a variable means.

Different Units of Measurement

→ When integrating two data sets, we have to pay attention to their respective units of measurement.

→ An example of this would be when we study the prices of gasoline in the world. To do this we gather data from different data providers. Data sets can contain prices per gallon and others can contain prices per liter.

Different Levels of Aggregation

→ Having different levels of aggregation is similar to having different types of measurement.

→ An example of this would be a data set containing data per week versus one containing data per work week.

→ This type of error is generally easy to detect, and *summarizing* (or the inverse, *expanding*) the data sets will fix it.

Step 4: Exploratory data analysis

→ During exploratory data analysis we take a deep dive into the data.

→ The goal of this step is to gain a deep understanding of the data.

→ Information becomes much easier to grasp when shown in a picture, therefore we mainly use graphical techniques to gain an understanding of data and the interactions between variables.

→ The main goal of this phase is discovering anomalies missed before, forcing to take a step back and fix them.

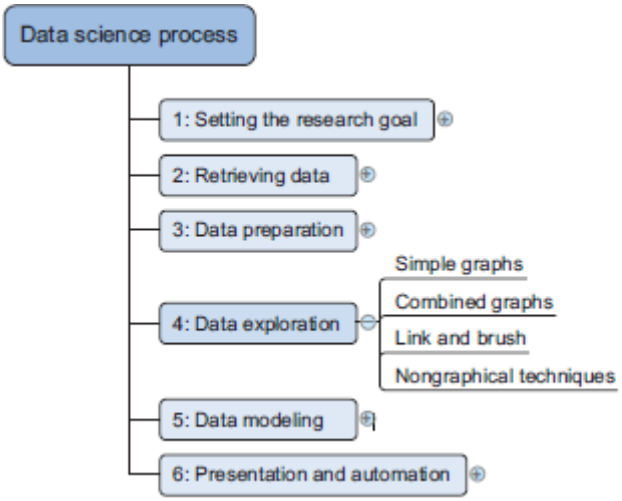


Figure 2.14 Step 4: Data exploration

→ The visualization techniques use in this phase range from simple line graphs or histograms to more complex diagrams such as Sankey and network graphs.

→ Sometimes it's useful to compose a composite graph from simple graphs to get even more insight into the data. Other times the graphs can be animated or made interactive to make it.

→ Another technique called *brushing and linking* can also be used. With brushing and linking we combine and link different graphs and tables (or views) so changes in one graph are automatically transferred to the other graphs. This interactive exploration of data facilitates the discovery of new insights.

→ Tabulation, clustering, and other modeling techniques can also be a part of exploratory analysis.

Step 5: Build the models

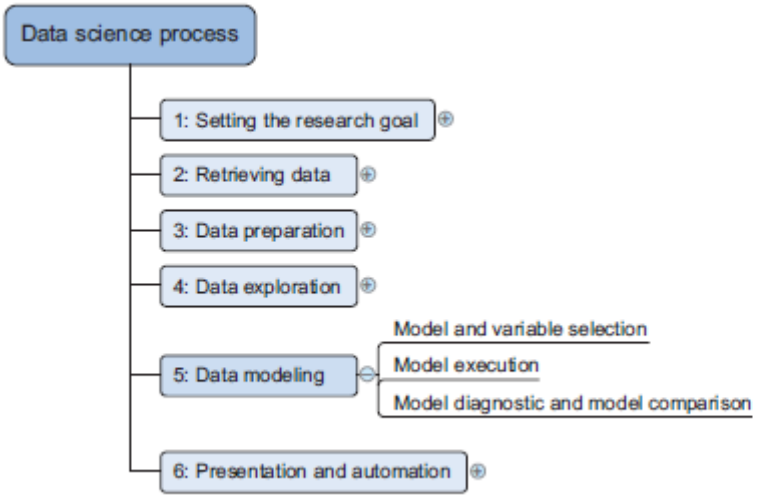


Figure 2.21 Step 5: Data modeling

→ Building a model is an iterative process.

→ Most models consist of the following main steps:

1. Selection of a modeling technique and variables to enter in the model
2. Execution of the model
3. Diagnosis and model comparison

Model and variable selection:

→ In this step it is need to select the variables we want to include in model and a modelling technique.

→ Exploratory analysis gives a fair idea of what variables will help to construct a good model.

→ Many modeling techniques are available, and choosing the right model for a problem requires judgment.

→ Model is selected based on

- model performance
- whether project meets all the requirements to model
- whether project is easy to implement
- the maintenance on the model
- whether the model easy to explain

Model execution:

→ This step involves implementing model in code.

→ Most programming languages, such as Python, already have libraries such as StatsModels or Scikit-learn. These packages use several of the most popular techniques.

→ Coding a model is a nontrivial task in most cases, so having these libraries available can speed up the process.

→ Only a handful of techniques have industry-ready implementations in Python. But it's fairly easy to use models that are available in R within Python with the help of the RPy library. RPy provides an interface from

Python to R. *R* is a free software environment, widely used for statistical computing.

Model diagnostics and model comparison

→ From multiple models we can choose the best one based on multiple criteria.

→ Working with a holdout sample helps to pick the best-performing model.

→ A holdout sample is a part of the data that leave out of the model building so it can be used to evaluate the model afterward.

→ The principle here is simple: the model should work on unseen data. use only a fraction of data to estimate the model and the other part, the holdout sample, is kept out of the equation.

→ The model is then unleashed on the unseen data and error measures are calculated to evaluate it. Multiple error measures are available.

→ Once the model is trained, calculate the model error with an error measure. Then we choose the model with the lowest error.

→ Many models make strong assumptions, such as independence of the inputs, and have to verify that these assumptions are indeed met. This is called *model diagnostics*.

Step 6: Presenting findings and building applications on top of them

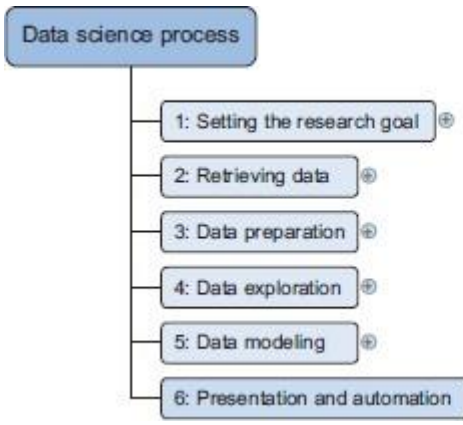


Figure 2.28 Step 6: Presentation and automation

→ The last step of the data science model is presenting results to the worldor stakeholders or business and automating the analysis, if needed.

→ The main aim of this phase is presenting the results to the stakeholders and automating or industrializing analysis process for repetitive reuse and integration with other tools

→ These results can take many forms, ranging from presentations to research reports.

→ Sometimes it is need to automate the execution of the process because the business will want to use the insights gained in another project or enable an operational process to use the outcome from model.

→ Certain projects require performing the business process over and over again, so automating the project will save time.

→ The last stage of the data science process is where *soft skills* will be most useful, and they're extremely important.