BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Data Science Process

Dr. Ramakrishna Dantu
Associate Professor, BITS Pilani

innovate achieve lead

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Introduction to Data Science

Data Science Process

- Data Science Methodology
  - Business understanding
  - Data Requirements
  - Data Acquisition
  - Data Understanding
  - Data preparation
  - Modelling
  - Model Evaluation
  - Deployment and feedback
- Case Study
- Data Science Proposal
  - Samples
  - Evaluation
  - Review Guide

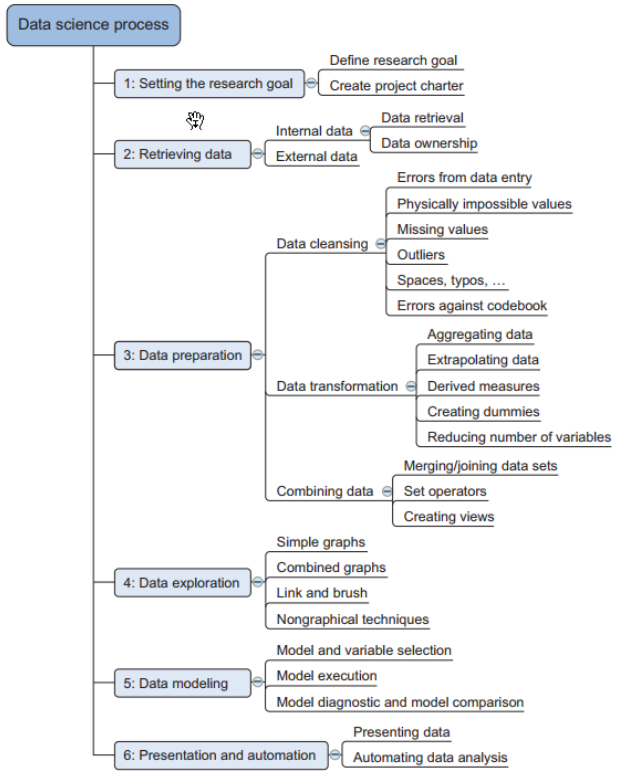
Data Science Process

Data Science Process

innovate achieve lead

Stages in Data Science Process

- Setting the research goal
- Retrieving data
- Data Preparation
- Data Exploration
- Data Modeling
- Presentation and Automation



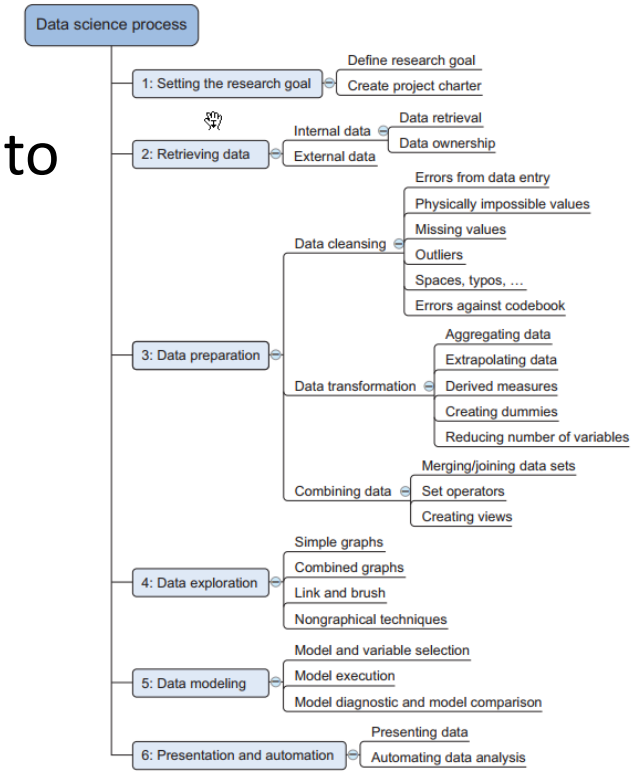
Source: Introducing Data Science by Cielin et al.,

Data Science Process

innovate achieve lead

Business Scenario

- Suppose you're working for a Bank
- The bank feels that it's losing too much money to bad loans and wants to reduce its losses
- To do so, they want a tool to help loan officers more accurately detect risky loans



Source: Introducing Data Science by Cielin et al.,

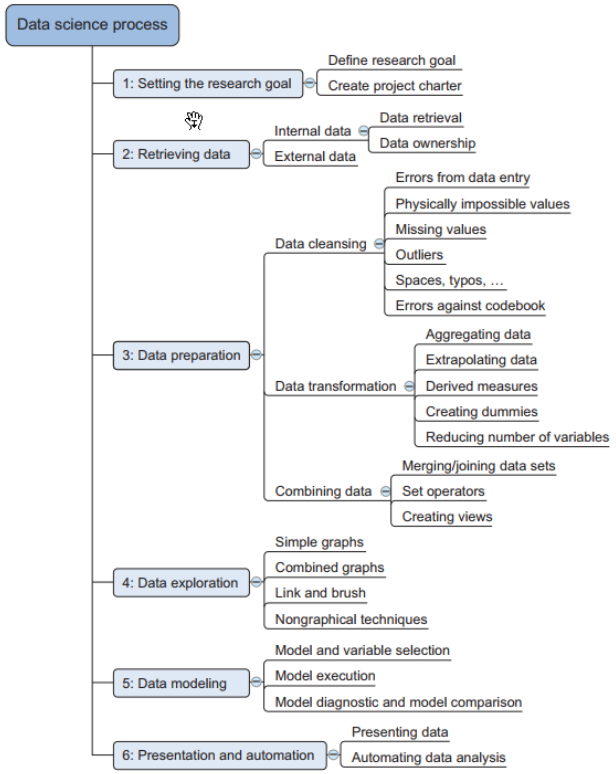
Setting the Research Goal

Data Science Process

innovate achieve lead

Setting the Research Goal - Concept

- An essential outcome of this stage is the research goal that states the purpose of your assignment in a clear and focused manner
- Understanding the business goals and context is critical for project success
- Ask questions until you grasp the exact business expectations
  - Identify how your project fits in the bigger picture
  - Understand how this research is going to change the business
  - Understand how business will use the results
- Nothing is more frustrating than when you report your findings back to the organization, everyone immediately realizes that you misunderstood their question
- Don't skim over this phase lightly
- Many data scientists fail here:
  - Despite their mathematical wit and scientific brilliance, they never seem to grasp the business goals and context



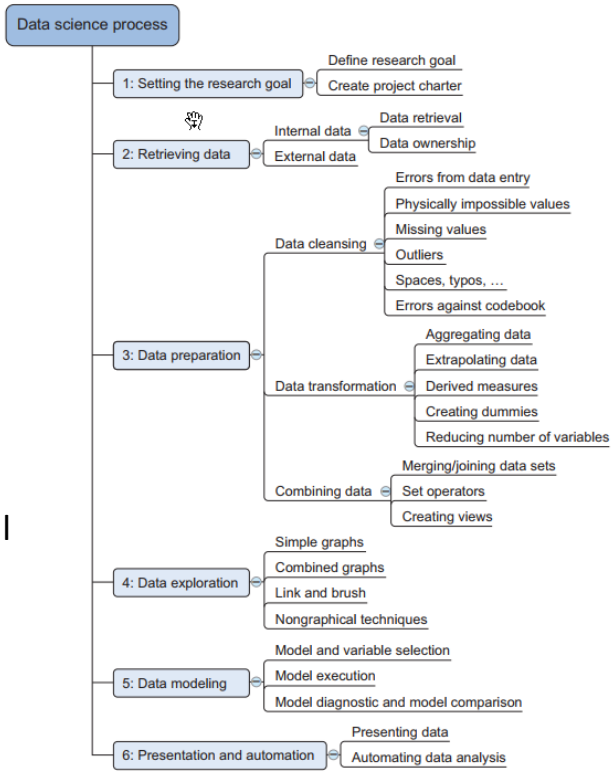
Source: Introducing Data Science by Cielin et al.,

Data Science Process

innovate achieve lead

Setting the Research Goal - Concept

- The first task in a data science project is to define a measurable and quantifiable goal
- Typically, this phase involves two steps:
  - Define research goal
  - Develop a project charter
- Our aim is to learn all about the context of the project:
  - Why do the sponsors want the project in the first place?
  - What do they lack, and what do they need?
  - What are they doing to solve the problem now, and why isn't that good enough?
  - What resources will you need: what kind of data and how much staff?
  - Will you have domain experts to collaborate with, and what are the computational
  - resources?
  - How do the project sponsors plan to deploy your results?
  - What are the constraints that have to be met for successful deployment?

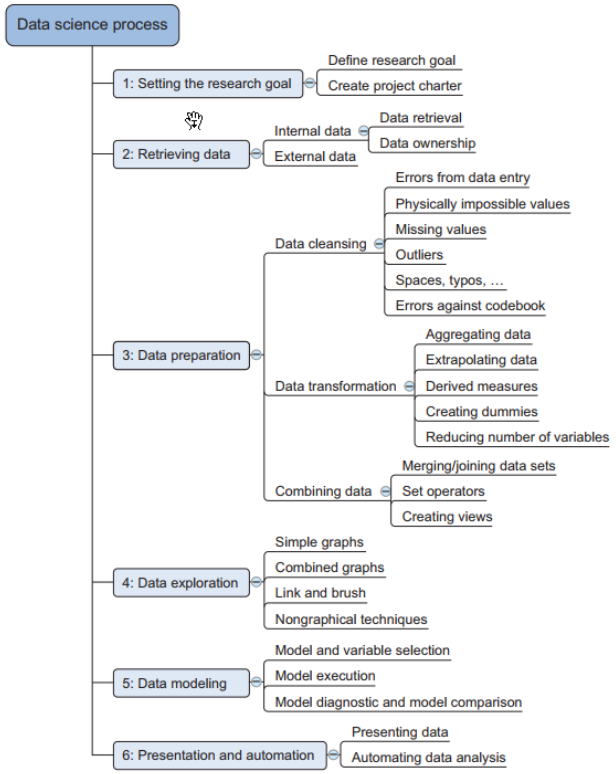


Source: Introducing Data Science by Cielin et al.,

Data Science Process

Setting the Research Goal - Concept

- A project charter requires teamwork, and your input covers at least the following:
  - A clear research goal
  - The project mission and context
  - How you're going to perform your analysis
  - What resources you expect to use
  - Proof that it's an achievable project, or proof of concepts
  - Deliverables and a measure of success
  - A timeline
- Your client can use this information to make an estimation of the project costs and the data and people required for your project to become a success.



Source: Introducing Data Science by Cielin et al.,

Data Science Process

Setting the Research Goal – Case Scenario

- In the loan example, the ultimate business goal is to reduce the bank's losses due to bad loans
- Project sponsor wants a tool to help loan officers more accurately score loan applicants, and so reduce the number of bad loans made
- Loan officers feel that they have final discretion on loan approvals
- Once we have a high-level understanding, we work with stakeholders to define the precise goal of the project
- The goal should be specific and measurable
  - NO - "We want to get better at finding bad loans" but instead
  - YES - "We want to reduce our rate of loan charge-offs by at least 10%, using a model that predicts which loan applicants are likely to default."

Source: Practical Data Science with R - Nina Zumel, John Mount

Data Science Process

innovate achieve lead

Setting the Research Goal – Case Scenario

- Importance of having specific and measureable goal
  - A concrete goal helps with acceptance criteria and when to stop the project
  - If the goal is less specific (not measurable), there is no boundary to the project, because no result will be "good enough."
  - If we don't know what we want to achieve, we don't know when to stop trying—or even what to try
  - When the project eventually terminates—because either time or resources run out—no one will be happy with the outcome

Source: Practical Data Science with R - Nina Zumel, John Mount

Data Science Process

Setting the Research Goal – Case Scenario

- When can we have non-specific or non-measurable goals?
  - When our project is explorative in nature
  - "Is there something in the data that correlates to higher defaults?"
  - "Should we think about reducing the kinds of loans we give out?"
    - Which types might we eliminate?
  - In this situation, we can still scope the project with concrete stopping conditions, such as a time limit
  - For example
    - We might decide to spend two weeks, and no more, exploring the data, with the goal of coming up with candidate hypotheses
    - These hypotheses can then be turned into concrete questions or goals for a full-scale modeling project

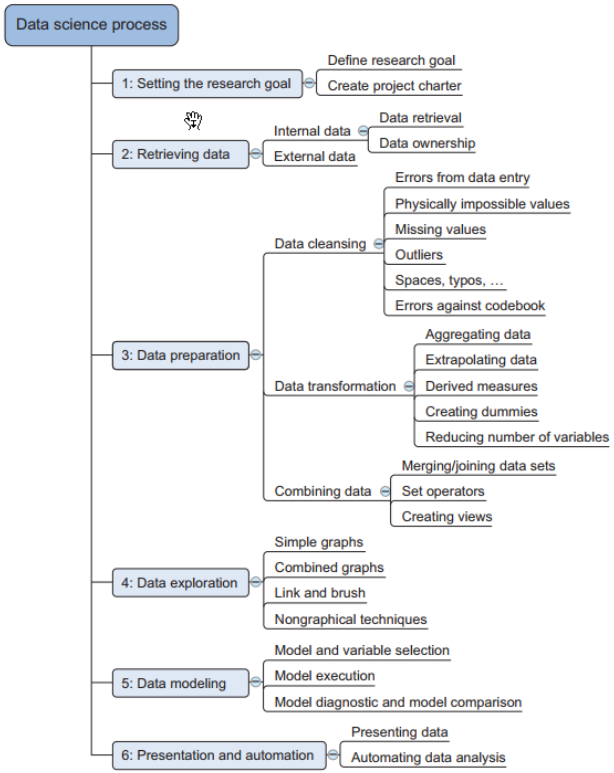
Source: Practical Data Science with R - Nina Zumel, John Mount

Data Collection

Data Science Process

Retrieving Data - Concept

- This step involves finding suitable data and getting access to it:
  - Internal data
  - External data
- In this step, we ask the following questions:
  - What data is available to me?
  - Will it help me solve the problem?
  - Is it enough?
  - Is the data quality good enough?
- Data found in this phase is usually in the raw form
- Data requires polishing and transformation before using it in the next stage



Source: Introducing Data Science by Cielin et al.,

Data Science Process

Retrieving Data – Case Scenario

- For the loan application case scenario, we need data related to loan applications, and those that were defaulted
- Imagine that we've collected a sample of representative loans from the last decade
- Some of the loans have defaulted
  - Most of them (about 70%) have not
- We've collected a variety of attributes about each loan application
  - See the listing in the next slide

Source: Practical Data Science with R - Nina Zumel, John Mount

Data Science Process

innovate achieve lead

Retrieving Data – Case Scenario

- Table 1.2 Loan data attributes

Status_of_existing_checking_account (at time of application)	Present_residence_since
Duration_in_month (loan length)	Collateral (car, property, and so on)
Credit_history	Age_in_years
Purpose (car loan, student loan, and so on)	Other_installment_plans (other loans/lines of credit—the type)
Credit_amount (loan amount)	Housing (own, rent, and so on)
Savings_Account_or_bonds (balance/amount)	Number_of_existing_credits_at_this_bank
Present_employment_since	Job (employment type)
Installment_rate_in_percentage_of_disposable_income	Number_of_dependents
Personal_status_and_sex	Telephone (do they have one)
Cosigners	Loan_status (dependent variable)

Source: Practical Data Science with R - Nina Zumel, John Mount

Data Science Process

Retrieving Data – Case Scenario

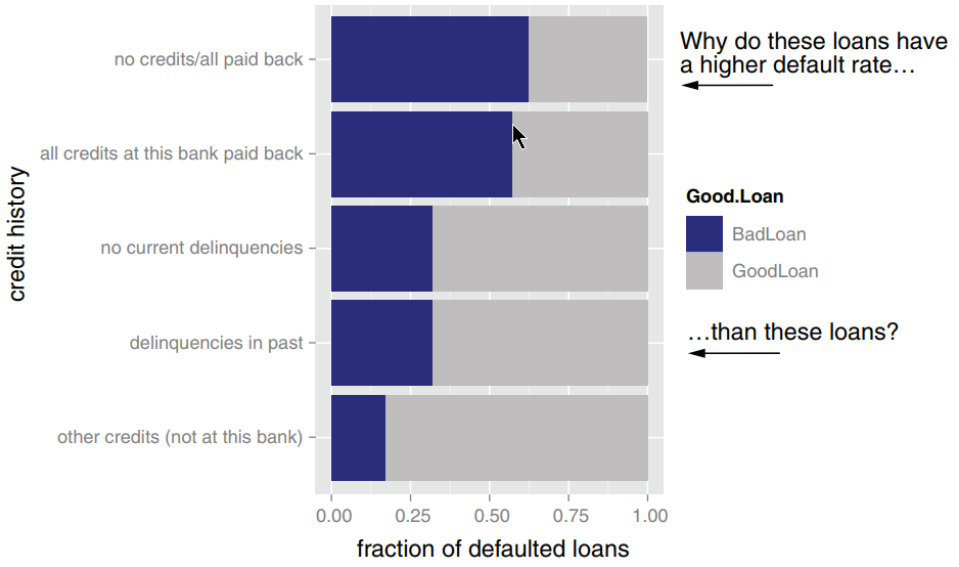
- In the dataset, Loan_status takes on two possible values:
  - GoodLoan and BadLoan
- For the purposes of this case scenario
  - GoodLoan means that is paid off, and a BadLoan means it is defaulted
- Analysis Tip:
  - As much as possible, try to use information that can be directly measured, rather than information that is inferred from another measurement
  - For example, we might be tempted to use income as a variable, reasoning that a lower income implies more difficulty paying off a loan
  - The ability to pay off a loan is more directly measured by considering the size of the loan payments relative to the borrower's disposable income
  - This information is more useful than income alone
  - This information is available in the dataset as the variable
    - Installment_rate_in_percentage_of_disposable_income

Source: Practical Data Science with R - Nina Zumel, John Mount

Data Science Process

Retrieving Data – Case Scenario

- This is the stage where we initially explore and visualize your data
- We also perform basic cleaning of the data, as needed
- In the process of exploring, we may discover that the data isn't suitable for your problem, or that you need other types of information as well
- We may discover things in the data that raise issues more important than the one you originally planned to address
- For example, the data in the figure seems counterintuitive.



The fraction of defaulting loans by credit history category
The dark region of each bar represents the fraction of loans in that category that defaulted.

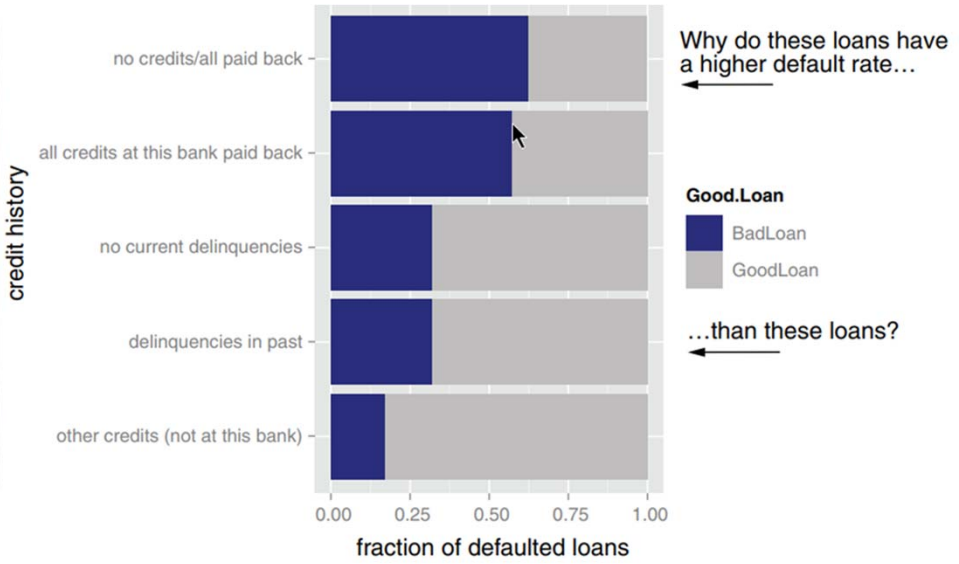
Source: Practical Data Science with R - Nina Zumel, John Mount

Data Science Process

innovate achieve lead

Retrieving Data – Case Scenario

- Why would some of the seemingly safe applicants (those who repaid all credits to the bank) default at a higher rate than seemingly riskier ones (those who had been delinquent in the past)?
- After looking more carefully and discussing with domain experts, we realize that this sample is inherently biased:
  - We only have loans that were actually made (and therefore already accepted).
- A true unbiased sample of loan applications should include
  - both loan applications that were accepted and ones that were rejected



Source: Practical Data Science with R - Nina Zumel, John Mount

Data Science Process

Retrieving Data – Case Scenario

- Because the data sample only includes accepted loans, there are fewer risky-looking loans than safe-looking ones in the data
- The probable story is that risky-looking loans were approved after a much stricter vetting process
  - A process that perhaps the safe-looking loan applications could bypass
- This suggests that if your model is to be used downstream of the current application approval process
  - Credit history is no longer a useful variable
- It also suggests that even seemingly safe loan applications should be more carefully scrutinized
- Discoveries like this may require us to approach stakeholders to change or refine the project goals
- We may decide to concentrate on the seemingly safe loan applications

Source: Practical Data Science with R - Nina Zumel, John Mount

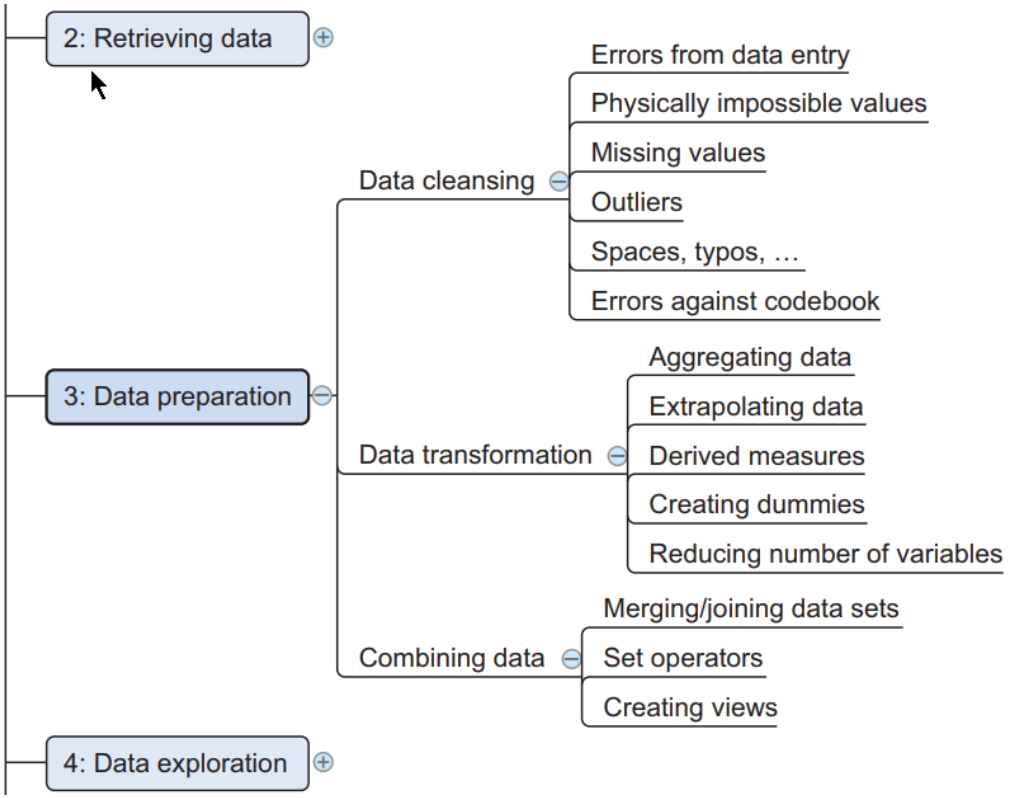
Data Preparation

Techniques used in Data Preparation

innovate achieve lead

Overview

- The raw data collected is likely to be "a diamond in the rough."
- Our task is to sanitize and prepare it for use in the modeling and reporting phase
- This includes cleansing, transformation, and combining the data from a raw form into data that's directly usable in your models
- Doing so is tremendously important because:
  - Our models will perform better with a clean data
  - Otherwise, garbage in equals garbage out



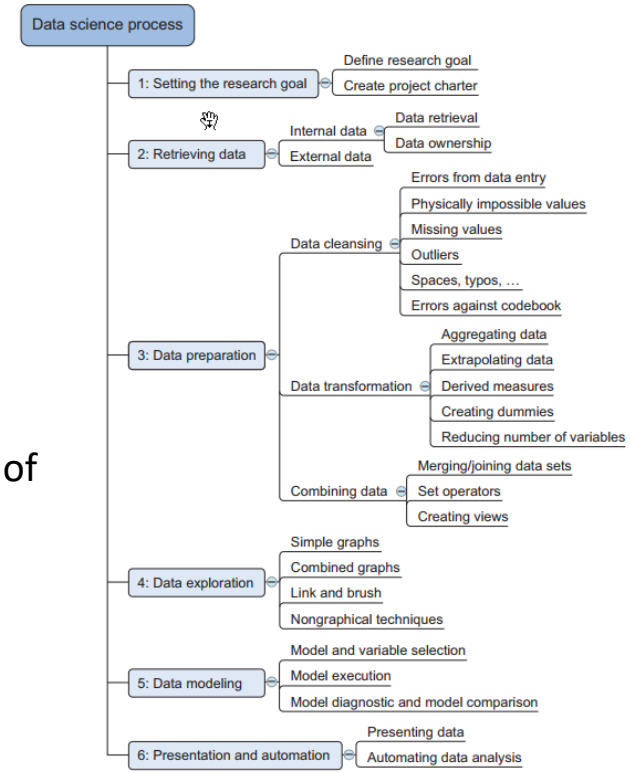
Source: Introducing Data Science by Cielin et al.,

Techniques used in Data Preparation

innovate achieve lead

Overview

- Data Preparation
  - The third phase of the process is data preparation
  - Cleansing
    - Finding and correcting errors, Spaces, Typos,
    - Missing values
    - Outliers
  - Transformation
    - Aggregating, Extrapolating, Derived measures, Reducing number of variables, Creating dummies, etc
  - Combining
    - Creating views
    - Set operators

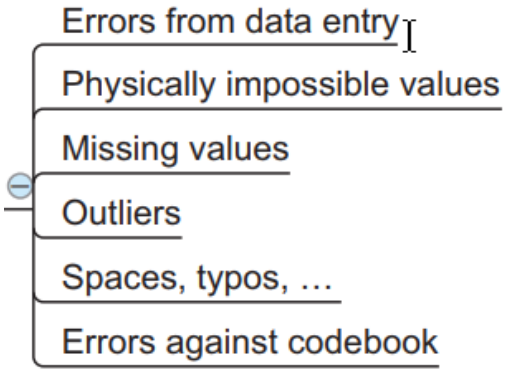


Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

Data Cleansing

- Data cleansing focuses on removing errors in the data
- The purpose is to give the data a "true and consistent representation" of the process from where it originates
- Two types of errors:
  - Interpretation error
    - It exists when we take the value in the data for granted
      - E.g, Age = 300 years, Weight = 500 kilos
  - Inconsistencies between data sources
    - E.g., Using "Male" is one table and "M" in another
    - E.g., Using "Rupees" in one table and "Dollars" in another



Source: Introducing Data Science by Cielin et al.,

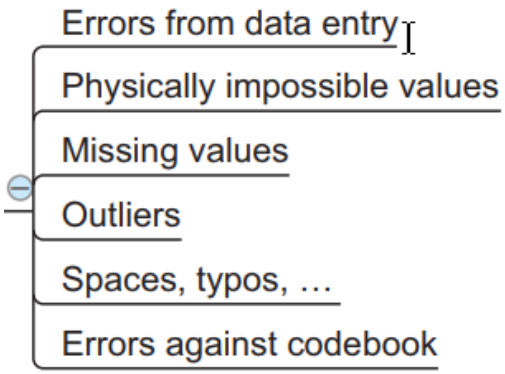
Techniques in Data Preparation

innovate achieve lead

Data Cleansing

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



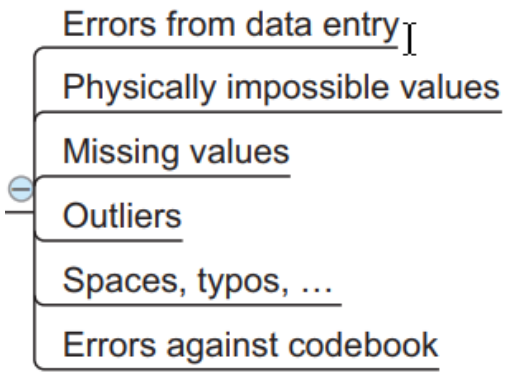
Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Cleansing

- Sometimes we may use more advanced methods, such as simple modeling, to find and identify data errors and anomalies
- Using diagnostic plots can be especially insightful
  - For example, simple regression can identify data points that seem out of place
- When a single observation has too much influence, this can point to an error in the data, but it can also be a valid point
- At the data cleansing stage, these advanced methods are, however, rarely applied and often regarded by certain data scientists as overkill.

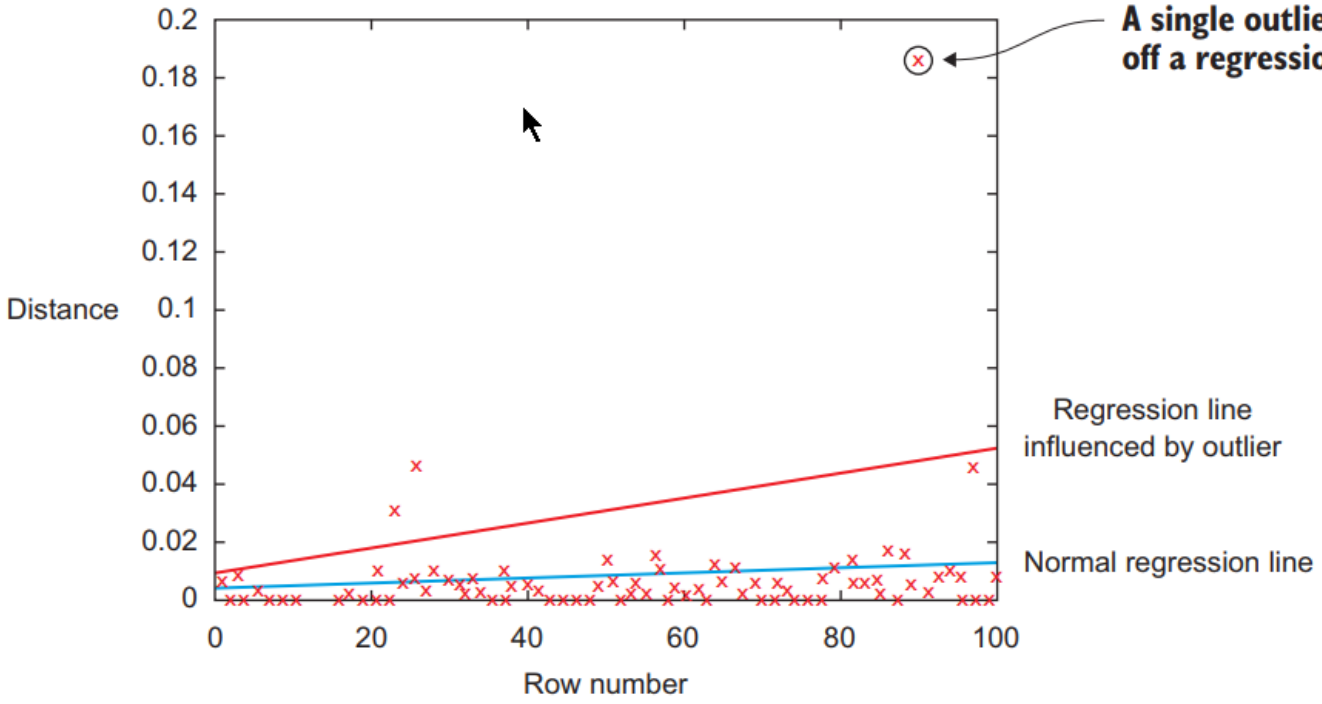


Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Cleansing



The outlier point influences the model heavily and is worth investigating because it can point to a region where we don't have enough data or might indicate an error in the data, but it also can be a valid data point

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Cleansing – Data Entry Errors

- Data entry errors may be caused by humans or machines
- Errors originating from machines are transmission errors or bugs in the extract, transform, and load phase (ETL)
- When the variable doesn't have many classes, error check can be done by tabulating the data with counts
- For example:
  - Consider the variable can take only two values: "Good" and "Bad"
  - We can create a frequency table and see if those are truly the only two values present
  - The values "Godo" and "Bade" point out something went wrong in at least 16 cases.

Table 2.3 Detecting outliers on simple variables with a frequency table

Value	Count
Good	1598647
Bad	1354468
Godo	15
Bade	1

Most errors of this type are easy to fix with simple assignment statements and if-then else rules:

```
if (x == "Godo"):
    x = "Good"
if (x == "Bade"):
    x = "Bad"
```

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Cleansing – Redundant White Space

- Whitespaces tend to be hard to detect but cause errors like other redundant characters would
- Most of us have lost several hours in projects because of a bug that was caused by whitespaces at the end of a string
- We require the software program to join two keys and notice that observations are missing from the output file
- After looking for days through the code, we realize that the cleaning during the ETL phase wasn't well executed, and keys in one table contained a whitespace at the end of a string
- This caused a mismatch of keys such as "FR " – "FR", dropping the observations that couldn't be matched
- Most programming languages provide string functions that will remove the leading and trailing whitespaces
- For instance, in Python provides the strip() function to remove leading and trailing spaces

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

Data Cleansing

- Fixing Capital Letter Mismatches
  - Capital letter mismatches are common.
  - Most programming languages make a distinction between "India" and "india"
  - We can solve this problem by applying a function that returns both strings in lowercase or uppercase, such as .lower() or .upper() in Python
    - "India".lower() == "india".lower() should result in true
- Impossible Values & Sanity Check
  - Sanity checks are another valuable type of data check
  - Here you check the value against physically or theoretically impossible values such as people taller than 3 meters or someone with an age of 300 years
  - Sanity checks can be directly expressed with rules:
    - ageCheck = 0 <= age <= 120

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Cleansing - Outliers

- An outlier is an observation that seems to be distant from other observations
- One observation that follows a different logic than the other observations
- The easiest way to find outliers is to use a plot or a table with the minimum and maximum values
- The plot on the top shows no outliers
- The plot on the bottom shows possible outliers on the upper side when a normal distribution is expected
- The normal distribution, or Gaussian distribution, is the most common distribution in natural sciences

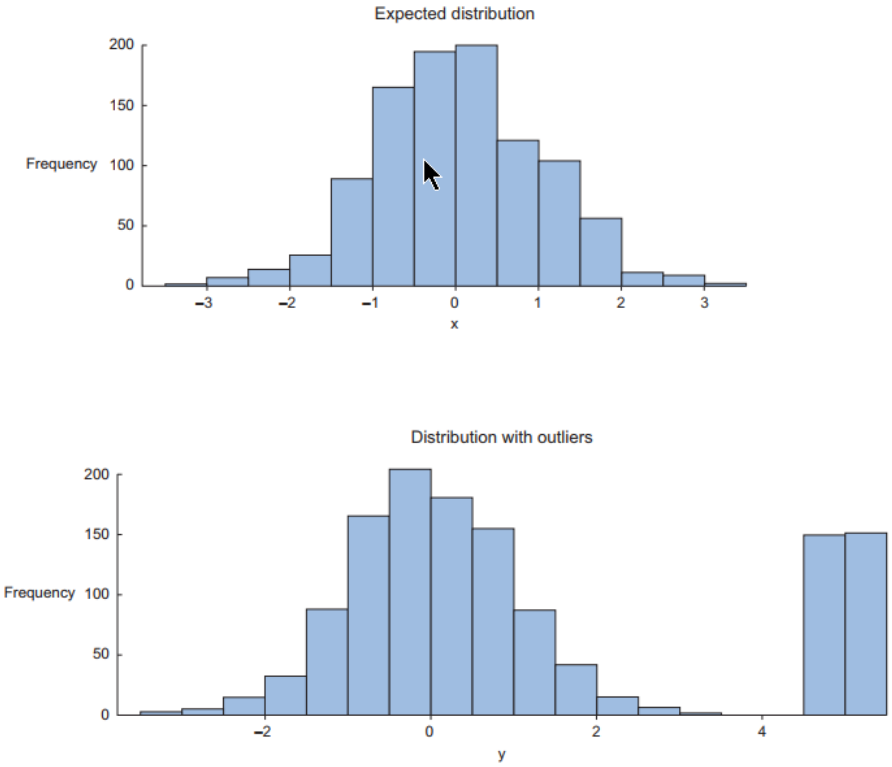


Figure 2.6 Distribution plots are helpful in detecting outliers and helping you understand the variable.

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Cleansing – Missing Values

- Missing values aren't necessarily wrong
- Missing values might be an indicator that
  - Something went wrong in the data collection or that an error happened in the ETL process
- Which technique to use at what time depends on a the particular case
- For instance, if we don't have observations to spare, omitting an observation is probably not an option
- If the variable can be described by a stable distribution, we could impute based on this
- However, a missing value may actually means "zero"?
- This can be the case in sales, for instance:
  - If no promotion is applied on a customer basket, that customer's promo is missing, but most likely it's also 0, no price cut

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

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

Data Cleansing

- Deviations from Code Book
  - A code book is a description of your 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
    - For example, "0" equals "negative", "5" stands for "very positive"
  - A code book also tells the type of data you're looking at:
    - Is it hierarchical, graph, something else?
  - You look at those values that are present in set A but not in set B
  - If we have multiple values to check, it's better to put them from the code book into a table and use a difference operator to check the discrepancy between both tables

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Cleansing

- Different Units of Measurement
  - When integrating two data sets, you have to pay attention to their respective units of measurement
  - For example, prices of gasoline in the world
  - When we gather data from different data providers, data sets can contain prices per gallon and others can contain prices per liter
  - A simple conversion will do the trick in this case
- Different Levels of Aggregation
  - Having different levels of aggregation is similar to having different types of measurement
  - For example, a data set containing data per week versus one containing data per month
  - This type of error is generally easy to detect, and summarizing (or the inverse, expanding) the data sets will fix it

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Transformation

- Relationships between an input variable and an output variable aren't always linear
- For instance, a relationship of the form $y = ae^{bx}$
- Taking the log of the independent variables simplifies the estimation problem dramatically
- Transforming the input variables greatly simplifies the estimation problem.
- Other times you might want to combine two variables into a new variable

x	1	2	3	4	5	6	7	8	9	10
log(x)	0.00	0.43	0.68	0.86	1.00	1.11	1.21	1.29	1.37	1.43
y	0.00	0.44	0.69	0.87	1.02	1.11	1.24	1.32	1.38	1.46

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Transformation

x	1	2	3	4	5	6	7	8	9	10
log(x)	0.00	0.43	0.68	0.86	1.00	1.11	1.21	1.29	1.37	1.43
y	0.00	0.44	0.69	0.87	1.02	1.11	1.24	1.32	1.38	1.46

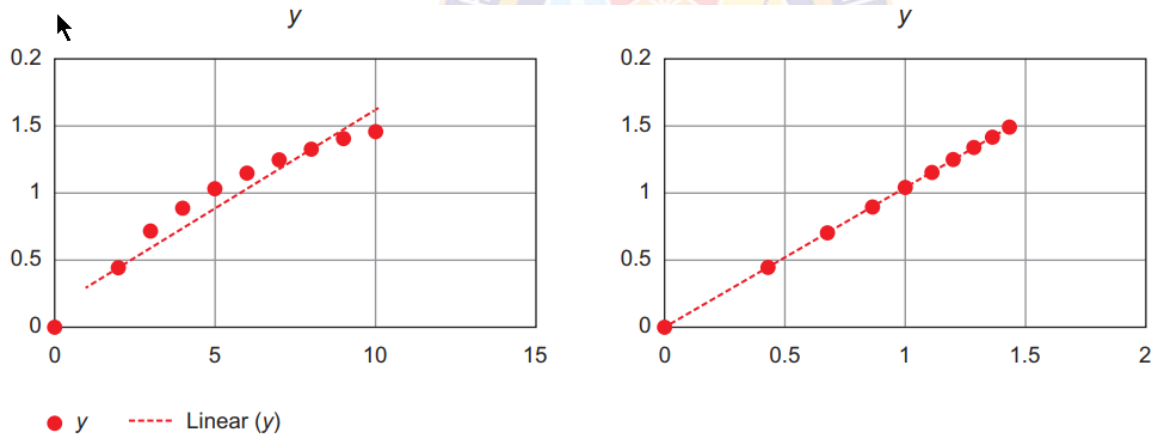


Figure 2.11 Transforming x to log x makes the relationship between x and y linear (right), compared with the non-log x (left).

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

Data Transformation

- Reducing the Number of Variables
  - Sometimes there are too many variables and need to reduce the number because they don't add new information to the model
  - Having too many variables in the model makes the model difficult to handle
  - Certain techniques don't perform well when you overload them with too many input variables
    - For instance, all the techniques based on a Euclidean distance perform well only up to 10 variables
  - Data scientists use special methods to reduce the number of variables but retain the maximum amount of data

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

Data Transformation

- Reducing the Number of Variables
  - Figure shows how reducing the number of variables makes it easier to understand the key values
  - It also shows how two variables account for 50.6% of the variation within the data set (component1 = 27.8% + component2 = 22.8%)
  - These variables, called "component1" and "component2," are both combinations of the original variables
  - They're the *principal components* of the underlying data structure

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Transformation

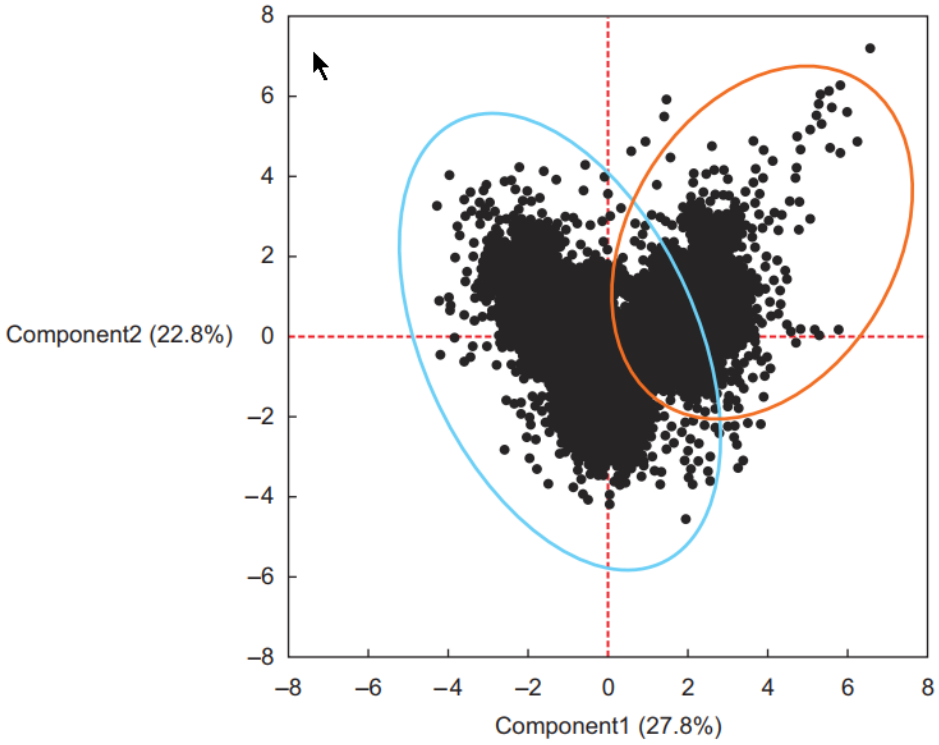


Figure 2.12 Variable reduction allows you to reduce the number of variables while maintaining as much information as possible.

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Transformation – Euclidean Distance

- Euclidean distance or "ordinary" distance is an extension to one of the first things we learn in trigonometry: Pythagoras's leg theorem
- If we know the length of the two sides next to the 90° angle of a right-angled triangle we can easily derive the length of the remaining side (hypotenuse)
- The formula for this is hypotenuse = $\sqrt{Side1^2 + Side2^2}$.
- The Euclidean distance between two points in a two-dimensional plane is calculated using a similar formula: distance = $\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$.
- If we want to expand this distance calculation to more dimensions, add the coordinates of the point within those higher dimensions to the formula
- For three dimensions we get distance = $\sqrt{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2}$
- Euclidean Distance is used in
  - K-nearest neighbors (classification) or K-means (clustering) to find the "k closest points" of a particular sample point
  - Hierarchical clustering, agglomerative clustering (complete and single linkage) where you want to find the distance between clusters.

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Data Transformation

- Turning Variables into Dummies
  - Variables can be turned into dummy variables
  - *Dummy variables* can only take two values:
    - true(1) or false(0)
  - They're used to indicate the absence of a categorical effect that may explain the observation
  - In this case you'll make separate columns for the classes stored in one variable and indicate it with 1 if the class is present and 0 otherwise
  - An example is turning one column named Weekdays into the columns Monday through Sunday
  - You use an indicator to show if the observation was on a Monday; you put 1 on Monday and 0 elsewhere

Customer	Year	Gender	Sales
1	2015	F	10
2	2015	M	8
1	2016	F	11
3	2016	M	12
4	2017	F	14
3	2017	M	13

M F

Customer	Year	Sales	Male	Female
1	2015	10	0	1
1	2016	11	0	1
2	2015	8	1	0
3	2016	12	1	0
3	2017	13	1	0
4	2017	14	0	1

Figure 2.13 Turning variables into dummies is a data transformation that breaks a variable that has multiple classes into multiple variables, each having only two possible values: 0 or 1.

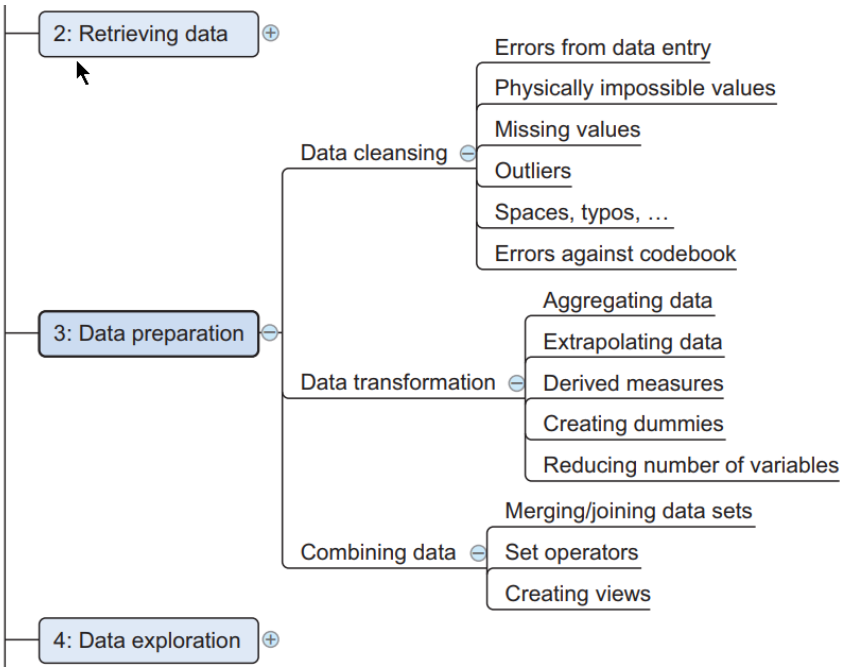
Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Combining Data

- Data comes from several different places
- In this substep we focus on integrating these different sources
- Data varies in size, type, and structure, ranging from databases and Excel files to text documents
- Different ways of combining data
  - Joining
    - Enriching an observation from one table with information from another table
  - Appending or Stacking
    - Adding the observations of one table to those of another table



Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Combining Data – Joining Tables

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Region
John Doe	NY
Jackie Qi	NC

Client	Item	Month	Region
John Doe	Coca-Cola	January	NY
Jackie Qi	Pepsi-Cola	January	NC

Figure 2.7 Joining two tables on the Item and Region keys

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Combining Data – Appending Tables

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Item	Month
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

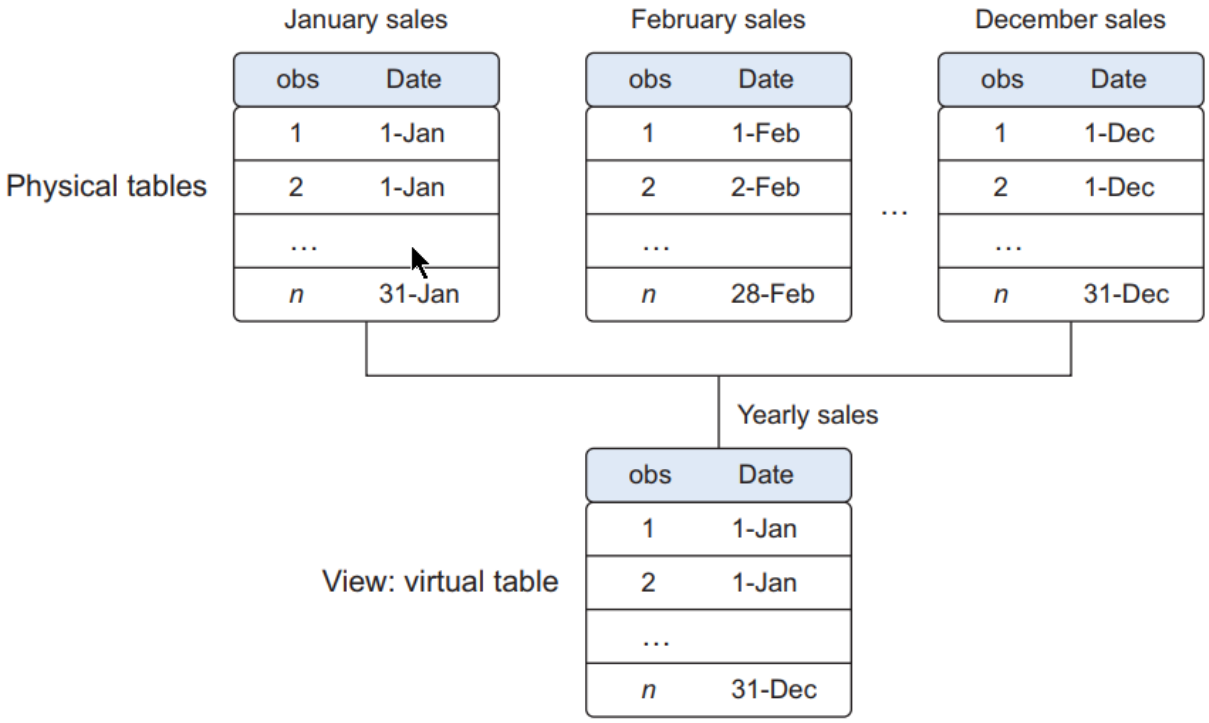
Figure 2.8 Appending data from tables is a common operation but requires an equal structure in the tables being appended.

Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Combining Data – Using Views to Simulate Joins & Appends



Source: Introducing Data Science by Cielin et al.,

Techniques in Data Preparation

innovate achieve lead

Combining Data – Enriching Aggregated Values

Product class	Product	Sales in $	Sales t-1 in $	Growth	Sales by product class	Rank sales
A	B	X	Y	(X-Y) / Y	AX	NX
Sport	Sport 1	95	98	–3.06%	215	2
Sport	Sport 2	120	132	–9.09%	215	1
Shoes	Shoes 1	10	6	66.67%	10	3

Figure 2.10 Growth, sales by product class, and rank sales are examples of derived and aggregate measures.

Source: Introducing Data Science by Cielin et al.,

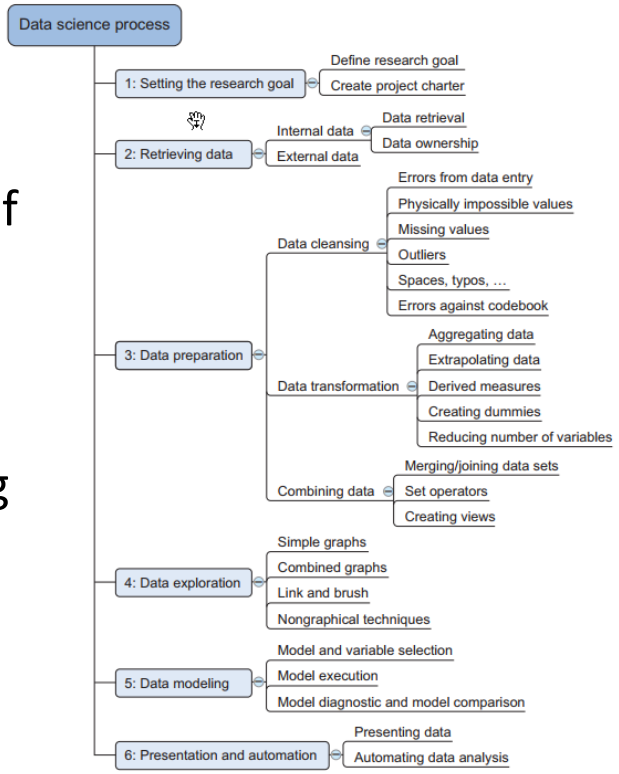
Data Exploration

Data Exploration

innovate achieve lead

Overview

- Data Exploration
  - The fourth phase in the process is data exploration
  - The goal of this step is to gain a deeper understanding of the data
  - We look for patterns, correlations, and deviations based on visual and descriptive techniques
  - The insights gained here will enable us to start modeling

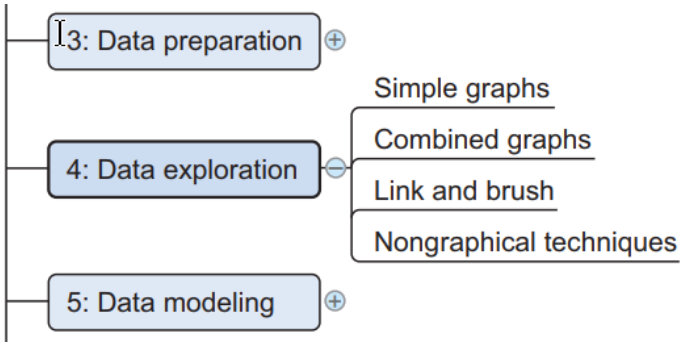


Source: Introducing Data Science by Cielin et al.,

Data Exploration

Overview

- During exploratory data analysis we take a deep dive into the data
- Information becomes much easier to grasp when shown in a picture
- We mainly use graphical techniques to gain an understanding of your data and the interactions between variables
- It's common to discover anomalies we missed previously
  - We may have to take a step back and fix them
- The visualization techniques in this phase can range from simple line graphs or histograms to more complex diagrams such as Sankey diagram (see next slide)



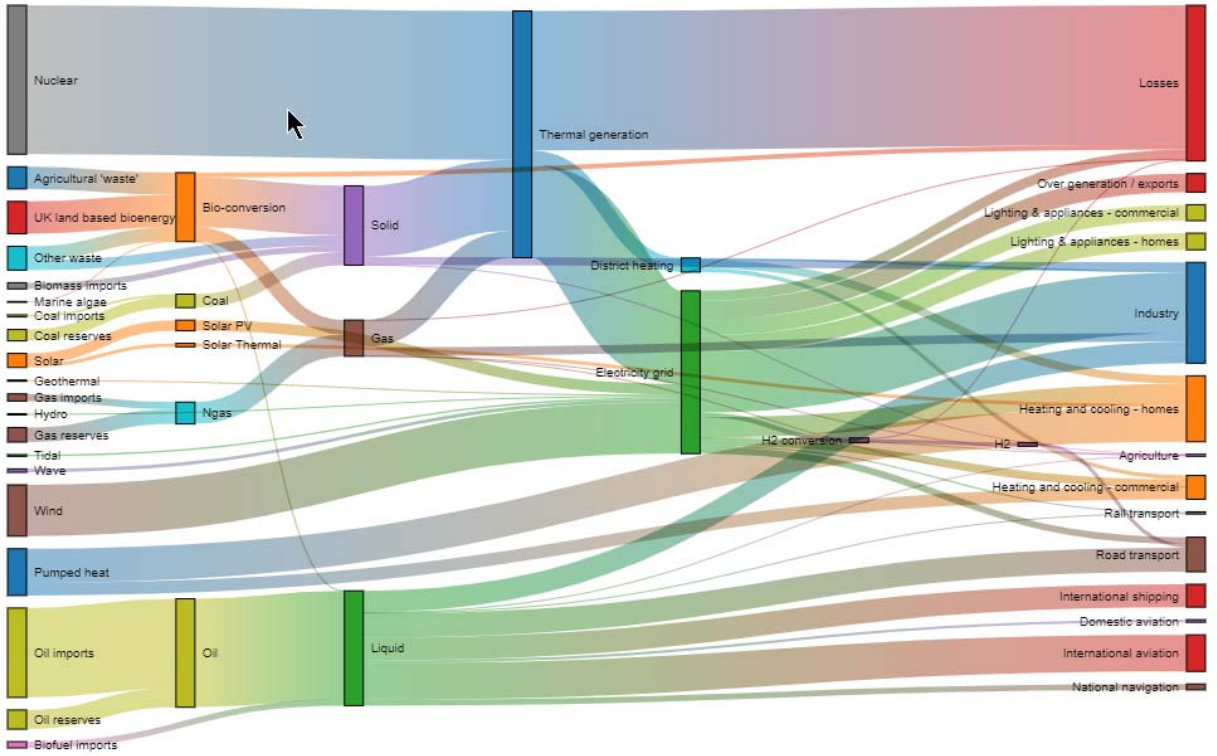
Source: Introducing Data Science by Cielin et al.,

Data Exploration

innovate achieve lead

Sankey diagram

- This Sankey diagram visualizes the flow of energy: supplies are on the left, and demands are on the right
- Links show how varying amounts of energy are converted or transmitted before being consumed or lost



Data: Department of Energy & Climate Change via Tom Counsell

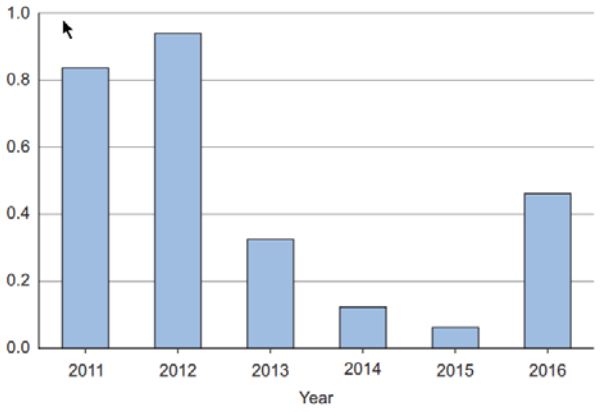
Source: https://observablehq.com/@d3/sankey-diagram

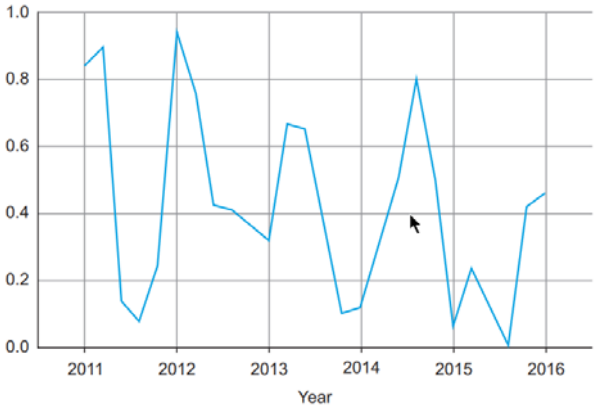
Data Exploration

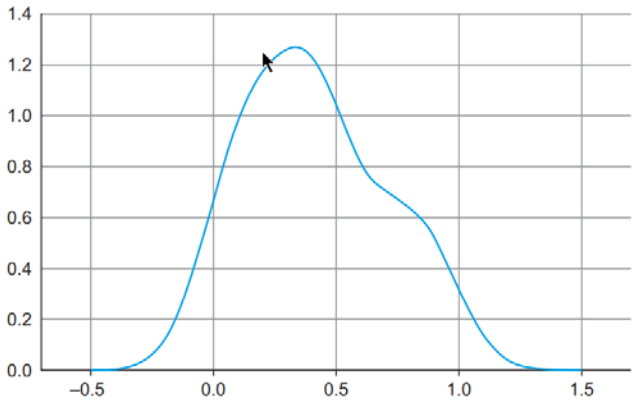
innovate achieve lead

Overview

- A bar chart, a line plot, and a distribution are some of the graphs used in exploratory analysis







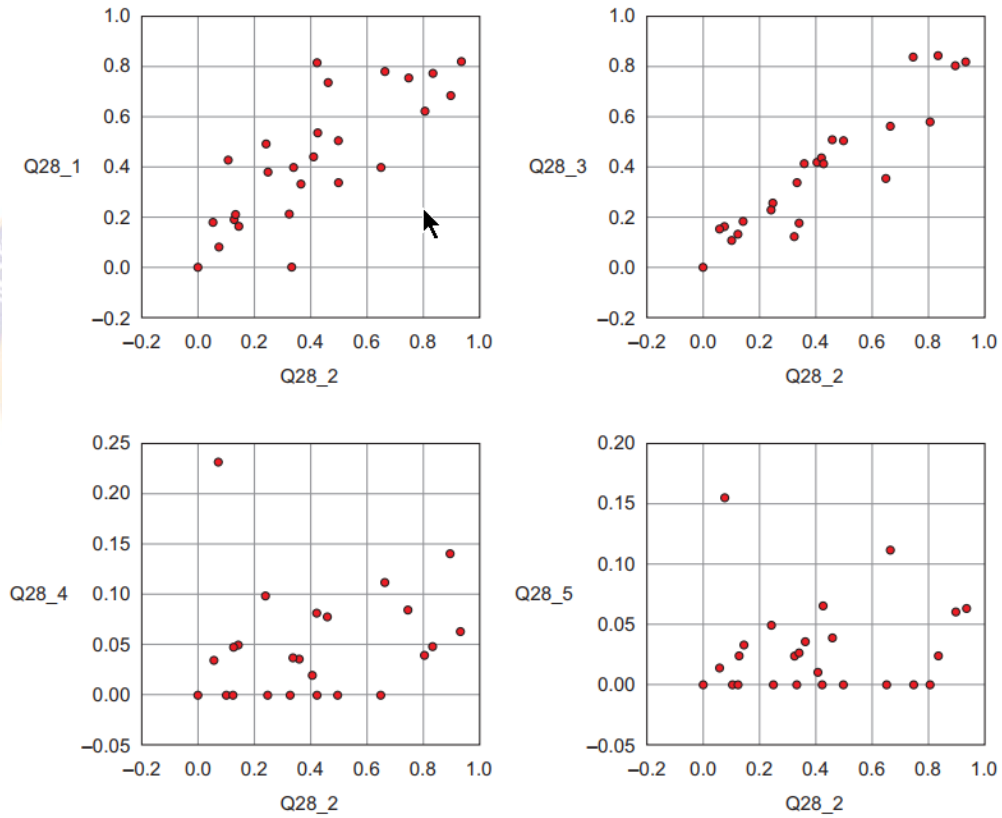
Source: Introducing Data Science by Cielin et al.,

Data Exploration

innovate achieve lead

Overview

- Drawing multiple plots together can help us understand the structure of your data over multiple variables.



Source: Introducing Data Science by Cielin et al.,

Data Exploration

innovate achieve lead

Overview

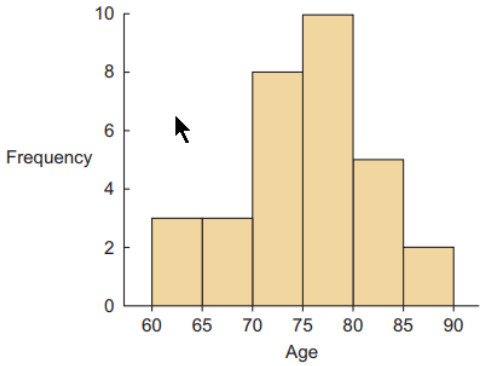


Figure 2.19 Example histogram: the number of people in the age-groups of 5-year intervals

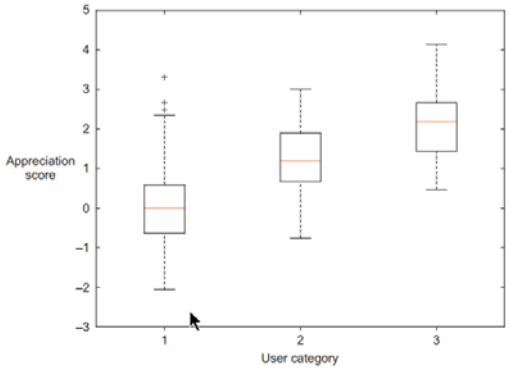


Figure 2.20 Example boxplot: each user category has a distribution of the appreciation each has for a certain picture on a photography website.

- Histogram
  - In a histogram a variable is cut into discrete categories and the number of occurrences in each category are summed up and shown in the graph
- Boxplot
  - The boxplot, on the other hand, doesn't show how many observations are present but does offer an impression of the distribution within categories
  - It can show the maximum, minimum, median, and other characterizing measures at the same time.

Source: Introducing Data Science by Cielin et al.,

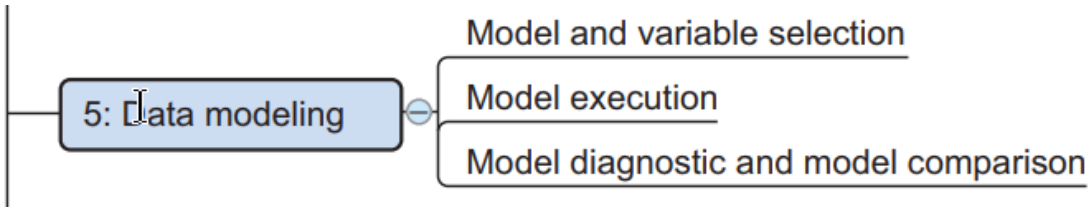
Data Modeling

Data Modeling

innovate achieve lead

Overview

- The most common data science modeling tasks are these:
  - Classifying—Deciding if something belongs to one category or another
  - Scoring—Predicting or estimating a numeric value, such as a price or probability
  - Ranking—Learning to order items by preferences
  - Clustering—Grouping items into most-similar groups
  - Finding relations—Finding correlations or potential causes of effects seen in the data
  - Characterizing—Very general plotting and report generation from data



Source: Practical Data Science with R - Nina Zumel, John Mount

Data Exploration

Modeling – Case Scenario

- The loan application problem is a classification problem:
  - We want to identify loan applicants who are likely to default
- Some common approaches in such cases are logistic regression and tree-based methods
- To solve this problem, we decide that a decision tree is most suitable
- Loan officers, who will use our model, are interested in knowing an indication of how confident the model is in its decision:
  - Is this applicant highly likely to default, or only somewhat likely?

Source: Practical Data Science with R - Nina Zumel, John Mount

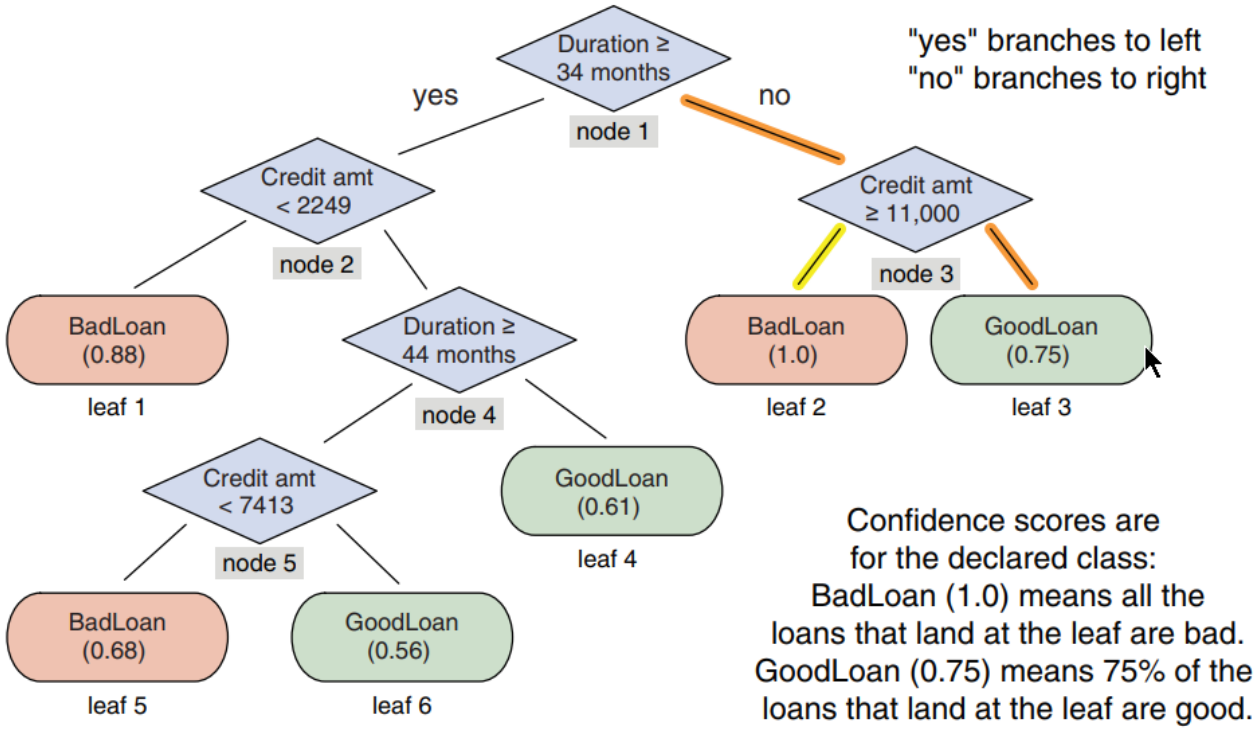
Data Exploration

innovate achieve lead

Modeling – Case Scenario

Let's suppose that there is an application for a one-year loan of $10,000

On the other hand, suppose that there is an application for a one-year loan of $15,000



Source: Practical Data Science with R - Nina Zumel, John Mount

Data Exploration

Modeling – Case Scenario

- Let's suppose that we discover the model shown in figure
- Let's trace an example path through the tree
- Let's suppose that there is an application for a one-year loan of $10,000
- At the top of the tree (node 1), the model checks if the loan is for longer than 34 months
- The answer is "no," so the model takes the right branch down the tree
  - This is shown as the highlighted branch from node 1
- The next question (node 3) is whether the loan is for more than DM 11,000
- Again, the answer is "no," so the model branches right and arrives at leaf 3.

Source: Practical Data Science with R - Nina Zumel, John Mount

Data Exploration

innovate achieve lead

Modeling – Case Scenario

- Historically, 75% of loans that arrive at this leaf are good loans, so the model recommends that you approve this loan, as there is a high probability that it will be paid off
- On the other hand, suppose that there is an application for a one-year loan of $15,000
- In this case, the model would first branch right at node 1, and then left at node 3, to arrive at leaf 2
- Historically, all loans that arrive at leaf 2 have defaulted, so the model recommends that you reject this loan application.

Source: Practical Data Science with R - Nina Zumel, John Mount


innovate
achieve
lead

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
ज्ञानं परमं बलम्

Thank You!