2 Collecting and Preparing Data

Figure 2.1 Periodic population surveys such as censuses help governments plan resources to supply public services. (credit: modification of work “Census 2010 @ La Fuente” by Jenn Turner/Flickr, CC BY 2.0)

Chapter Outline

Introduction

Data collection and preparation are the first steps in the data science cycle. They involve systematically gathering the necessary data to meet a project's objectives and ensuring its readiness for further analysis. Well-executed data collection and preparation serve as a solid foundation for effective, data-driven decisionmaking and aid in detecting patterns, trends, and insights that can drive business growth and efficiency.

With today’s ever-increasing volume of data, a robustapproach to data collection is crucial for ensuring accurate and meaningful results. This process requires following a comprehensive and systematic methodology designed to ensure the quality, reliability, and validity of data gathered for analysis. It involves identifying and sourcing relevant data from diverse sources, including internal databases, external repositories, websites, and user-generated information. And it requires meticulous planning and execution to guarantee the accuracy, comprehensiveness, and reliability of the collected data.

Preparing, or “wrangling,” the collected data adequately prior to analysis is equally important. Preparation involves scrubbing, organizing, and transforming the data into a format suitable for analysis. Data preparation plays a pivotal role in detecting and resolving any inconsistencies or errors present in the data, thereby enabling accurate analysis. The rapidly advancing technology and widespread use of the internet have added complexity to the data collection and preparation processes. As a result, data analysts and organizations face many challenges, such as identifying relevant data sources, managing large data volumes, identifying outliers or erroneous data, and handling unstructured data. By mastering the art and science of collecting and preparing data, organizations can leverage valuable insights to drive informed decision-making and achieve business success.

2.1 Overview of Data Collection Methods

Learning Outcomes

By the end of this section, you should be able to:

- 2.1.1 Define data collection and its role in data science.
- 2.1.2 Describe different data collection methods commonly used in data science, such as surveys and experiments.
- 2.1.3 Recognize scenarios where specific data collection methods are most appropriate.

Data collection refers to the systematic and well-organized process of gathering and accurately conveying important information and aspects related to a specific phenomenon or event. This involves using statistical tools and techniques to collect data, identify its attributes, and capture relevant contextual information. The gathered data is crucial for making sound interpretations and gaining meaningful insights. Additionally, it is important to take note of the environment and geographic location from where the data was obtained, as it can significantly influence the decision-making process and overall conclusions drawn from the data.

Data collection can be carried out through various methods, depending on the nature of the research or project and the type of data being collected. Some common methods for data collection include experiments, surveys, observation, focus groups, interviews, and document analysis.

This chapter will focus on the use of surveys and experiments to collect data. Social scientists, marketing specialists, and political analysts regularly use surveys to gather data on topics such as public opinion, customer satisfaction, and demographic information. Pharmaceutical companies heavily rely on experimental data from clinical trials to test the safety and efficacy of new drugs. This data is then used by their legal teams to gain regulatory approval and bring drugs to market.

Before collecting data, it is essential for a data scientist to have a clear understanding of the project's objectives, which involves identifying the research question or problem and defining the target population or sample. If a survey or experiment is used, the design of the survey/experiment is also a critical step, requiring careful consideration of the type of questions, response options, and overall structure. A survey may be conducted online, via phone, or in person, while experimental research requires a controlled environment to ensure data validity and reliability.

Types of Data

Observational and transactional data play important roles in data analysis and related decision-making, each offering unique insights into different aspects of real-world phenomena and business operations.

Observational data, often used in qualitative research, is collected by systematically observing and recording behavior without the active participation of the researcher. Transactional data refers to any type of information related to transactions or interactions between individuals, businesses, or systems, and it is more often used in quantitative research.

Many fields of study use observational data for their research. Table 2.1 summarizes some examples of fields that rely on observational data, the type of data they collect, and the purpose of their data collection.

Field	Data Collected By	Purpose
Education	Teachers	To monitor and assess student behavior and learning progress in the classroom
Psychology	Therapists and psychologists	To gather information about their clients' behavior, thoughts, and emotions
Health care	Medical professionals	To diagnose and monitor patients' conditions and progress
Market research	Businesses	To gather information about consumer behavior and preferences to improve their products, services, and marketing strategies

Table 2.1 Fields Where Observation Methods Are Used

Environmental science	Scientists	To gather data about the natural environment and track changes over time
Criminal investigations	Law enforcement officers	To gather evidence and information about criminal activity
Animal behavior	Zoologists	To study and understand the behavior of various animal species
Transportation planning	Urban planners and engineers	To collect data on traffic patterns and transportation usage to make informed decisions about infrastructure and transit systems

Table 2.1 Fields Where Observation Methods Are Used

Transactional data is collected by directly recording transactions that occur in a particular setting, such as a retail store or an online platform that allows for accurate and detailed information on actual consumer behavior. It can include financial data, but it also includes data related to customer purchases, website clicks, user interactions, or any other type of activity that is recorded and tracked.

Transactional data can be used to understand patterns and trends, make predictions and recommendations, and identify potential opportunities or areas for improvement. For example, the health care industry may focus on transactional data related to patient interactions with health care providers and facilities, such as appointments, treatments, and medications prescribed. The retail industry may use transactional data on customer purchases and product returns, while the transportation industry may analyze data related to ticket sales and passenger traffic.

While observational data provides detailed descriptions of behavior, transactional data provides numerical data for statistical analysis. There are strengths and limitations with each of these, and the examples in this chapter will make use of both types.

EXAMPLE 2.1

Problem

Ashley loves setting up a bird feeder in her backyard and watching the different types of birds that come to feed. She has always been curious about the typical number of birds that visit her feeder each day and has estimated the number based on the amount of food consumed. However, she has to visit her grandmother's house for three days and is worried about leaving the birds without enough food. In order to prepare the right amount of bird food for her absence, Ashley has decided to measure the total amount of feed eaten each day to determine the total amount of food needed for her three-day absence. Which method of data collection is best suited for Ashley's research on determining the total amount of food required for her three-day absence—observational or transactional? Provide a step-by-step explanation of the chosen method.

Solution

Ashley wants to ensure that there is enough food for her local birds while she is away for three days. To do this, she will carefully observe the feeder daily for two consecutive weeks. She will record the total amount of feed eaten each day and make sure to refill the feeder each morning before the observation. This will provide a consistent amount of food available for the birds. After two weeks, Ashley will use the total amount of food consumed and divide it by the number of days observed to estimate the required daily food. Then, she will multiply the daily food by three to determine the total amount of bird food needed for her three-day absence. By directly observing and recording the bird food, as well as collecting data for two weeks, Ashley will gather accurate and reliable information. This will help her confidently prepare the necessary amount of bird food for her feathered friends while she is away, thus ensuring that the birds are well-fed and taken care of during her absence.

EXAMPLE 2.2

Problem

A group of data scientists working for a large hospital have been tasked with analyzing their transactional data to identify areas for improvement. In the past year, the hospital has seen an increase in patient complaints about long wait times for appointments and difficulties scheduling follow-up visits. Samantha is one of the data scientists tasked to collect data in order to analyze these issues.

a. What methodology should be employed by Samantha to collect pertinent data for analyzing the recent surge in patient complaints regarding extended appointment wait times and difficulties in scheduling follow-up visits at the hospital?

b. What strategies could be used to analyze the data?

Solution

a. Explore the stored information as transactional data

b. Collecting transactional data for analysis can be achieved by utilizing various sources within the hospital setting. These sources include:

1. Electronic Health Records (EHRs): Samantha can gather data from the hospital's electronic health records system. This data may include patients' appointment schedules, visit durations, and wait times. This information can help identify patterns and trends in appointment scheduling and wait times.

2. Appointment Booking System: Samantha can gather data from the hospital's appointment booking system. This data can include appointment wait times,

appointment types (e.g., primary care, specialist), and scheduling difficulties (e.g., appointment availability, cancellations). This information can help identify areas where the booking system may be causing delays or challenges for patients.

3. Hospital Call Center: Samantha can gather data from the hospital's call center, which is responsible for booking appointments over the phone. This data can include call wait times, call duration, and reasons for call escalations. This information can help identify areas for improvement in the call center's processes and procedures.

4. Historical Data: Samantha can analyze historical data, such as appointment wait times and scheduling patterns, to identify any changes that may have contributed to the recent increase in complaints. This data can also be compared to current data to track progress and improvements in wait times and scheduling.

Collecting Data Through Experiments

Collecting data through scientific experiments requires a well-designed experimental scheme, describing the research objectives, variables, and procedures. The establishment of a control specimen is crucial, and data is obtained through systematic properties, measurements, or characteristics. It is crucial to follow ethical guidelines for the proper documentation and ethical utilization of the collected data (see Ethics in Data Collection).

Consider this example: Scientist Sally aimed to investigate the impact of sunlight on plant growth. The research inquiry was to determine whether increased exposure to sunlight enhances the growth of plants. Sally experimented with two groups of plants wherein one group received eight hours of sunlight per day, while the other only received four hours. The height of each plant was measured and documented every week for four consecutive weeks. The main research objective was to determine the growth rate of plants exposed to eight hours of sunlight compared to those with only four hours. A total of 20 identical potted plants were used, with one group allocated to the "sunlight" condition and the other to the "limited sunlight" condition. Both groups were maintained under identical environmental conditions, including temperature, humidity, and soil moisture. Adequate watering was provided to ensure equal hydration of all plants. The measurements of plant height were obtained and accurately recorded every week. This approach allowed for the collection of precise and reliable data on the impact of sunlight on plant growth, which can serve as a valuable resource for further research and understanding of this relationship.

2.2 Survey Design and Implementation

Learning Outcomes

By the end of this section, you should be able to:

- 2.2.1 Describe the elements of survey design and identify the steps data scientists take to ensure the reliability of survey results.
- 2.2.2 Describe methods for avoiding bias in survey questions.
- 2.2.3 Describe various sampling techniques and the advantages of each.

Surveys are a common strategy for gathering data in a wide range of domains, including market research, social sciences, and education. Surveys collect information from a sampleof individuals and often use questionnaires to collect data. Sampling is the process of selecting a subset of a larger population to represent and analyze information about that population.

Designing the Survey

The process of data collection through surveys is a crucial aspect of research—and one that requires careful planning and execution to gather accurate and reliable data. The first step, as stated earlier, is to clearly define the research objectives and determine the appropriate target population. This will help you structure the survey and identify the specific questions that need to be included.

Constructing good surveys is hard. A survey should begin with simple and easy-to-answer questions and progress to more complex or sensitive ones. This can help build a rapport with the respondents and increase their willingness to answer more difficult questions. Additionally, the researcher may consider mixing up the response options for multiple-choice questions to avoid response bias. To ensure the quality of the data collected, the survey questionnaire should undergo a pilot test with a small group of individuals from the target population. This allows the researcher to identify any potential issues or confusion with the questions and make necessary adjustments before administering the survey to the larger population.

Open-Ended Versus Closed-Ended Questions

Surveys should generally contain a mix of closed-ended and open-ended questions to gather both quantitative and qualitative data.

Open-ended questions allow for more in-depth responses and provide the opportunity for unexpected insights. They also allow respondents to elaborate on their thoughts and provide detailed and personal responses. Closed-ended questions have predetermined answer choices and are effective in gathering quantitative data. They are quick and easy to answer, and their clear and structured format allows for quantifiable results.

Avoiding Bias in Survey Questions

Unbiased sampling and unbiased survey methodology are essential for ensuring accurate and reliable results. One well-known real-life instance of sampling bias leading to inaccurate findings is the 1936 Literary Digest poll. This survey aimed to forecast the results of the US presidential election and utilized a mailing list of telephone and automobile owners. This approach was considered biased toward affluent individuals and therefore favored Republican voters. As a consequence, the poll predicted a victory for Republican nominee Alf Landon. However, the actual outcome was a landslide win for Franklin D. Roosevelt (Lusinchi, 2012). This discrepancy can be attributed to the biased sampling method as well as the use of primarily closed-ended questions, which may not have accurately captured the opinions of all voters.

An example of a biased survey question in a survey conducted by a shampoo company might be "Do you prefer our brand of shampoo over cheaper alternatives?" This question is biased because it assumes that the respondent prefers the company's brand over others. A more unbiased and accurate question would be "What factors do you consider when choosing a shampoo brand?" This allows for a more detailed and accurate response. The biased question could have led to inflated results in favor of the company's brand.

Sampling

The next step in the data collection process is to choose a participant sampleto ideally represent the restaurant's customer base. Sampling could be achieved by randomly selecting customers, using customer databases, or targeting specific demographics, such as age or location.

Sampling is necessary in a wide range of data science projects to make data collection more manageable and cost-effective while still drawing meaningful conclusions. A variety of techniques can be employed to determine a subset of data from a larger population to perform research or construct hypotheses about the entire population. The choice of a sampling technique depends upon the nature and features of the population being studied as well as the objectives of the research. When using a survey, researchers must also consider the tool(s) that will be used for distributing the survey, such as through email, social media, or physically distributing questionnaires at the restaurant. It's crucial to make the survey easily accessible to the chosen sample to achieve a higher response rate.

A number of sampling techniques and their advantages are described below. The most frequently used among these are simple random selection, stratified sampling, cluster sampling, and convenience sampling.

1. Simple random selection. Simple random selection is a statistical technique used to pick a representative sample from a larger population. This process involves randomly choosing individuals or items from the population, ensuring that each selected member of the population has an identical chance of being contained in the sample. The main step in simple random selection is to define the population of interest and assign a unique identification number to each member. This could be done using a random number generator, a computer program designed to generate a sequence of random numbers, or a random number table, which lists numbers in a random sequence. The primary benefit of this technique is its ability to minimize bias and deliver a fair representation of the population.

In the health care field, simple random sampling is utilized to select patients for medical trials or surveys, allowing for a diverse and unbiased sample (Elfil & Negida, 2017). Similarly, in finance, simple random sampling can be applied to gather data on consumer behavior and guide decision-making in financial institutions. In engineering, this technique is used to select random samples of materials or components for quality control testing. In the political arena, simple random sampling is commonly used to select randomly registered voters for polls or surveys, ensuring equal representation and minimizing bias in the data collected.

2. Stratified sampling. Stratified sampling involves splitting the population into subgroups based on specified factors, such as age, area, income, or education level, and taking a random sample from each stratum in proportion to its size in the population. Stratified sampling allows for a more accurate representation of the population as it ensures that all subgroups are adequately represented in the sample. This can be especially useful when the variables being studied vary significantly between the stratified groups.

3. Cluster sampling. With cluster sampling, the population is divided into natural groups or clusters, such as schools, communities, or cities, with a random sample of these clusters picked and all members within the chosen clusters included in the sample. Cluster sampling is helpful to represent the entire population even if it is difficult or time-consuming due to challenges such as identifying clusters, sourcing a list of clusters, traveling to different clusters, and communicating with them. Additionally, data analysis and sample size calculation may be more complex, and there is a risk of bias in the sample. However, cluster sampling can be more cost-effective.

An example of cluster sampling would be a study on the effectiveness of a new educational program in a state. The state is divided into clusters based on school districts. The researcher uses a random selection process to choose a sample of school districts and then collects data from all the schools within those districts. This method allows the researcher to obtain a representative sample of the state's student population without having to visit each individual school, saving time and resources.

4. Convenience sampling. Convenience sampling applies to selecting people or items for the sample based on their availability and convenience to the data science research. For example, a researcher may choose to survey students in their classroom or manipulate data from social media users. Convenience sampling is effortless to achieve, and it is useful for exploratory studies. However, it may not provide a representative sample as it is prone to selection bias in that individuals who are more readily available or willing to participate may be overrepresented.

An example of convenience sampling would be conducting a survey about a new grocery store in a busy shopping mall. A researcher stands in front of the store and approaches people who are coming out of the store to ask them about their shopping experience. The researcher only includes responses from those who agreed to participate, resulting in a sample that is convenient but may not be representative of the entire population of shoppers in the mall.

5. Systematic sampling. Systematic sampling is based on starting at a random location in the dataset and then selecting every nth member from a population to be contained in the sample. This process is straightforward to implement, and it provides a representative sample when the population is randomly distributed. However, if there is a pattern in the sampling frame (the organizing structure that represents the population from which a sample is drawn), it may lead to a biased sample.

Suppose a researcher wants to study the dietary habits of students in a high school. The researcher has a list of all the students enrolled in the school, which is approximately 1,000 students. Instead of randomly selecting a sample of students, the researcher decides to use systematic sampling. The researcher first assigns a number to each student, going from 1 to 1,000. Then, the researcher randomly selects a number from 1 to 10—let's say they select 4. This number will be the starting point for selecting the sample of students. The researcher will then select every 10th student from the list, which means every student with a number ending in 4 (14, 24, 34, etc.) will be included in the sample. This way, the researcher will have a representative sample of 100 students from the high school, which is 10% of the population. The sample will consist of students from different grades, genders, and backgrounds, making it a diverse and representative sample.

6. Purposive sampling. With purposive sampling, one or more specific criteria are used to select participants who are likely to provide the most relevant and useful information for the research study. This can involve selecting participants based on their expertise, characteristics, experiences, or behaviors that are relevant to the research question.

For example, if a researcher is conducting a study on the effects of exercise on mental health, they may use purposive sampling to select participants who have a strong interest or experience in physical fitness and have a history of mental health issues. This sampling technique allows the researcher to target a specific population that is most relevant to the research question, making the findings more applicable and generalizable to that particular group. The main advantage of purposive sampling is that it can save time and resources by focusing on individuals who are most likely to provide valuable insights and information. However, researchers need to be transparent about their sampling strategy and potential biases that may arise from purposely selecting certain individuals.

7. Snowball sampling. Snowball sampling is typically used in situations where it is difficult to access a particular population; it relies on the assumption that people with similar characteristics or experiences tend to associate with each other and can provide valuable referrals. This type of sampling can be useful in studying hard-to-reach or sensitive populations, but it may also be biased and limit the generalizability of findings.

8. Quota sampling. Quota sampling is a non-probability sampling technique in which experimenters select participants based on predetermined quotas to guarantee that a

certain number or percentage of the population of interest is represented in the sample. These quotas are based on specific demographic characteristics, such as age, gender, ethnicity, and occupation, which are believed to have a direct or indirect relationship with the research topic. Quota sampling is generally used in market research and opinion polls, as it allows for a fast and cost-effective way to gather data from a diverse range of individuals. However, it is important to note that the results of quota sampling may not accurately represent the entire population, as the sample is not randomly selected and may be biased toward certain characteristics. Therefore, the findings from studies using quota sampling should be interpreted with caution.

9. Volunteer sampling. Volunteer sampling refers to the fact that the participants are not picked at random by the researcher, but instead volunteer themselves to be a part of the study. This type of sampling is commonly used in studies that involve recruiting participants from a specific population, such as a specific community or organization. It is also often used in studies where convenience and accessibility are important factors, as participants may be more likely to volunteer if the study is easily accessible to them. Volunteer sampling is not considered a random or representative sampling technique, as the participants may not accurately represent the larger population. Therefore, the results obtained from volunteer sampling may not be generalizable to the entire population.

Sampling Error

Sampling error is the difference between the results obtained from a sample and the true value of the population parameter it is intended to represent. It is caused by chance and is inherent in any sampling method. The goal of researchers is to minimize sampling errors and increase the accuracy of the results. To avoid sampling error, researchers can increase sample size, use probability sampling methods, control for extraneous variables, use multiple modes of data collection, and pay careful attention to question formulation.

Sampling Bias

Sampling bias occurs when the sample used in a study isn't representative of the population it intends to generalize to, leading to skewed or inaccurate conclusions. This bias can take many forms, such as selection

bias, where certain groups are systematically over- or underrepresented, or volunteer bias, where only a

specific subset of the population participates. Researchers use the sampling techniques summarized earlier to avoid sampling bias and ensure that each member of the population has an equal chance of being included in the sample. Additionally, careful consideration of the sampling frame should ideally encompass all members of the target population and provide a clear and accessible way to identify and select individuals or units for inclusion in the sample. Sampling bias can occur at various stages of the sampling process, and it can greatly impact the accuracy and validity of research findings.

Measurement Error

Measurement errors are inaccuracies or discrepancies that surface during the process of collecting, recording, or analyzing data. They may occur due to human error, environmental factors, or inherent inconsistencies in the phenomena being studied. Random error, which arises unpredictably, can affect the precision of measurements, and systematic errormay consistently bias measurements in a particular direction. In data analysis, addressing measurement error is crucial for ensuring the reliability and validity of results. Techniques for mitigating measurement error include improving data collection methods, calibrating instruments, conducting validation studies, and employing statistical methods like error modeling or sensitivity analysis to account for and minimize the impact of measurement inaccuracies on the analysis outcomes.

A Sampling Case Study

Consider a research study that wants to randomly select a group of college students from a larger population to examine the effects of exercise on their mental health outcomes. Using student ID numbers generated by a computer program, 100 participants from the larger population were randomly selected to participate in the study to achieve the desired accuracy. This process ensured that every student in the university had an equal chance of being selected to participate. The participants were then randomly assigned to either the exercise group or the control group. This method of random sampling ensures that the sample is representative of the larger population, providing a more accurate representation of the relationship between exercise and mental health outcomes for college students.

Types of sampling error that could occur in this study include the following:

1. Sampling bias. One potential source of bias in this study is self-selection bias. As the participants are all college students, they may not be representative of the larger population, as college students tend to have more access and motivation to exercise compared to the general population. This could limit the generalizability of the study's findings. In addition, if the researchers only recruit participants from one university, there may be under-coverage bias. This means that certain groups of individuals, such as nonstudents or students from other universities, may be excluded from the study, potentially leading to biased results.

2. Measurement error. Measurement errors could occur, particularly if the researchers are measuring the participants' exercise and mental health outcomes through self-

report measures. Participants may not accurately report their exercise habits or mental health symptoms, leading to inaccurate data.

3. Non-response bias. Some participants in the study may choose not to participate or may drop out before the study is completed. This could introduce non-response bias, as those who choose not to participate or drop out may differ from those who remain in the study in terms of their exercise habits or mental health outcomes.

4. Sampling variability. The sample of 100 participants is a relatively small subset of the larger population. As a result, there may be sampling variability, meaning that the characteristics and outcomes of the participants may differ from those of the larger population simply due to chance.

5. Sampling error in random assignment. In this study, the researchers randomly assign participants to either the exercise group or the control group. However, there is always a possibility of sampling error in the random assignment process, meaning that the groups may not be perfectly balanced in terms of their exercise habits or other characteristics.

These types of sampling errors can affect the accuracy and generalizability of the study's findings.

Researchers need to be aware of these potential errors and take steps to minimize them when designing and conducting their studies.

EXAMPLE 2.3

Problem

Mark is a data scientist who works for a marketing research company. He has been tasked to lead a study to understand consumer behavior toward a new product that is about to be launched in the market. As data scientists, they know the importance of using the right sampling technique to collect accurate and reliable data. Mark divided the population into different groups based on factors such as age, education, and income. This ensures that he gets a representative sample from each group, providing a more accurate understanding of consumer behavior. What is the name of the sampling technique used by Mark to ensure a representative sample from different groups of consumers for his study on consumer behavior toward a new product?

Solution

The sampling technique used by Mark is called stratified sampling. This involves dividing the population into subgroups or strata based on certain characteristics and then randomly selecting participants from each subgroup. This ensures that each subgroup is represented in the sample, providing a more accurate representation of the entire population. This type of sampling is often used in market research studies to get a more comprehensive understanding of consumer behavior and preferences. By using stratified sampling, Mark can make more reliable conclusions and recommendations for the new product launch based on the data he collects.

2.3 Web Scraping and Social Media Data Collection Learning Outcomes:

By the end of this section, you should be able to:

- 2.3.1 Discuss the uses of web scraping for collecting and preparing data for analysis.
- 2.3.2 Apply regular expressions for data manipulation and pattern matching.
- 2.3.3 Write Python code to scrape data from the web.
- 2.3.4 Apply various methods for parsing, extracting, processing, and storing data.

Web scraping and social media data collection are two approaches used to gather data from the internet. Web scraping involves pulling information and data from websites using a web data extraction tool, often known as a web scraper. One example would be a travel company looking to gather information about hotel prices and availability from different booking websites. Web scraping can be used to automatically gather this data from the various websites and create a comprehensive list for the company to use in its business strategy without the need for manual work.

Social media data collection involves gathering information from various platforms like Twitter and Instagram using application programming interface or monitoring tools. An application programming interface (API) is a set of protocols, tools, and definitions for building software applications allowing different software systems to communicate and interact with each other and enabling developers to access data and services from other applications, operating systems, or platforms. Both web scraping and social media data collection require determining the data to be collected and analyzing it for accuracy and relevance.

Web Scraping

There are several techniques and approaches for scraping data from websites. See Table 2.2 for some of the common techniques used. (Note: The techniques used for web scraping will vary depending on the website and the type of data being collected. It may require a combination of different techniques to effectively scrape data from a website.)

Web Scraping Technique	Details
Web Crawling	• Follows links on a web page to navigate to other pages and collect data from them • Useful for scraping data from multiple pages of a website
XPath	• Powerful query language • Navigates through the elements in an HTML document • Often used in combination with HTML parsing to select specific elements to scrape
Regular Expressions	• Search for and extract specific patterns of text from a web page • Useful for scraping data that follows a particular format, such as dates, phone numbers, or email addresses
HTML Parsing	• Analyzes the HTML (HyperText Markup Language) structure of a web page and identifies the specific tags and elements that contain the desired data • Often used for simple scraping tasks
Application Programming Interfaces (APIs)	• Authorize developers to access and retrieve data instantly without the need for web scraping • Often a more efficient and reliable method for data collection
(XML) API Subset	• XML (Extensible Markup Language) is another markup language used exchanging data • This method works similarly to using the HTML API subset by making HTTP requests to the website's API endpoints and then parsing the data received in XML format

(JSON) API Subset	• JSON (JavaScript Object Notation) is a lightweight data interchange format that is commonly used for sending and receiving data between servers and web applications • Many websites provide APIs in the form of JSON, making it another efficient method for scraping data

Table 2.2 Techniques and Approaches for Scraping Data from Websites

Social Media Data Collection

Social media data collection can be carried out through various methods such as API integration, social listening, social media surveys, network analysis, and image and video analysis. APIs provided by social media platforms allow data scientists to collect structured data on user interactions and content. Social listening involves monitoring online conversations for insights on customer behavior and trends. Surveys conducted on social media can provide information on customer preferences and opinions. Network analysis, or the examination of relationships and connections between users, data, or entities within a network, can reveal influential users and communities. It involves identifying and analyzing influential individuals or groups as well as understanding patterns and trends within the network. Image and video analysis can provide insights into visual trends and user behavior.

An example of social media data collection is conducting a Twitter survey on customer satisfaction for a food delivery company. Data scientists can use Twitter's API to collect tweets containing specific hashtags related to the company and analyze them to understand customers' opinions and preferences. They can also use social listening to monitor conversations and identify trends in customer behavior. Additionally, creating a social media survey on Twitter can provide more targeted insights into customer satisfaction and preferences. This data can then be analyzed using data science techniques to identify key areas for improvement and drive informed business decisions.

Using Python to Scrape Data from the Web

As noted previously, web scraping is a strategy of gathering data from the internet using automated mechanisms or programs. Python is one of the popular programming languages used for web scraping due to its various libraries and frameworks that make it easy to pull and process data from websites.

To scrape data such as a table from a website using Python, we follow these steps:

1. Import the pandas library. The first step is to import the pandas library (https://openstax.org/r/pandas), which is a popular Python library for data analysis and manipulation.

import pandas as pd

2. Use the read_html**()** function. This function is used to read HTML tables from a web page and convert them into a list of DataFrame objects. Recall from What Are Data and Data Science? that a DataFrame is a data type that pandas uses to store multi-column tabular data. df = pd.read_html("https://......

3. Access the desired data. If the data on the web page is divided into different tables, we need to specify which table we want to extract. We have used indexing to access the desired table (for example: index 4) from the list of DataFrame objects returned by the read_html() function. The index here represents the table order in the web page.

4. Store the data in a DataFrame. The result of the read_html() function is a list of DataFrame objects, and each DataFrame represents a table from the web page. We can store the desired data in a DataFrame variable for further analysis and manipulation.

5. Display the DataFrame. By accessing the DataFrame variable, we can see the extracted data in a tabular format.

6. Convert strings to numbers. As noted in Chapter 1, a string is a data type used to represent a sequence of characters, such as letters, numbers, and symbols that are enclosed by matching single (') or double (") quotes. If the data in the table is in string format and we want to perform any numerical operations on it, we need to convert the data to numerical format. We can use the to_numeric() function from pandas to convert strings to numbers and then store the result in a new column in the DataFrame. df['column_name'] = pd.to_numeric(df['column_name'])

This will create a new column in the DataFrame with the converted numerical values, which can then be used for further analysis or visualization.

In computer programming, indexing usually starts from 0. This is because most programming languages use 0 as the initial index for arrays, matrices, or other data structures. This convention has been adopted to simplify the implementation of some algorithms and to make it easier for programmers to access and manipulate data. Additionally, it aligns with the way computers store and access data in memory. In the context of parsing tables from HTML pages, using 0 as the initial index allows programmers to easily access and manipulate data from different tables on the same web page. This enables efficient data processing and analysis, making the task more manageable and

EXAMPLE2.4

Proble

Extract data table "Current Population Survey: Household Data: (Table A-13). Employed an persons by occupation, Not seasonally adjusted" FRED (Federal Reserve Economic (https://openstax.org/r/fred) website in the link https://fred.stlouisfed.org/release/tables?rid=50&eid=3149#snid=4498 (https://openstax.org/r/stlouisfed) using Python code. The data in table provides a representation of the overall employment and unemployment situation ir States. The table is organized into two main sections: employed persons and unemploye

Solutio

PYTHON CODE

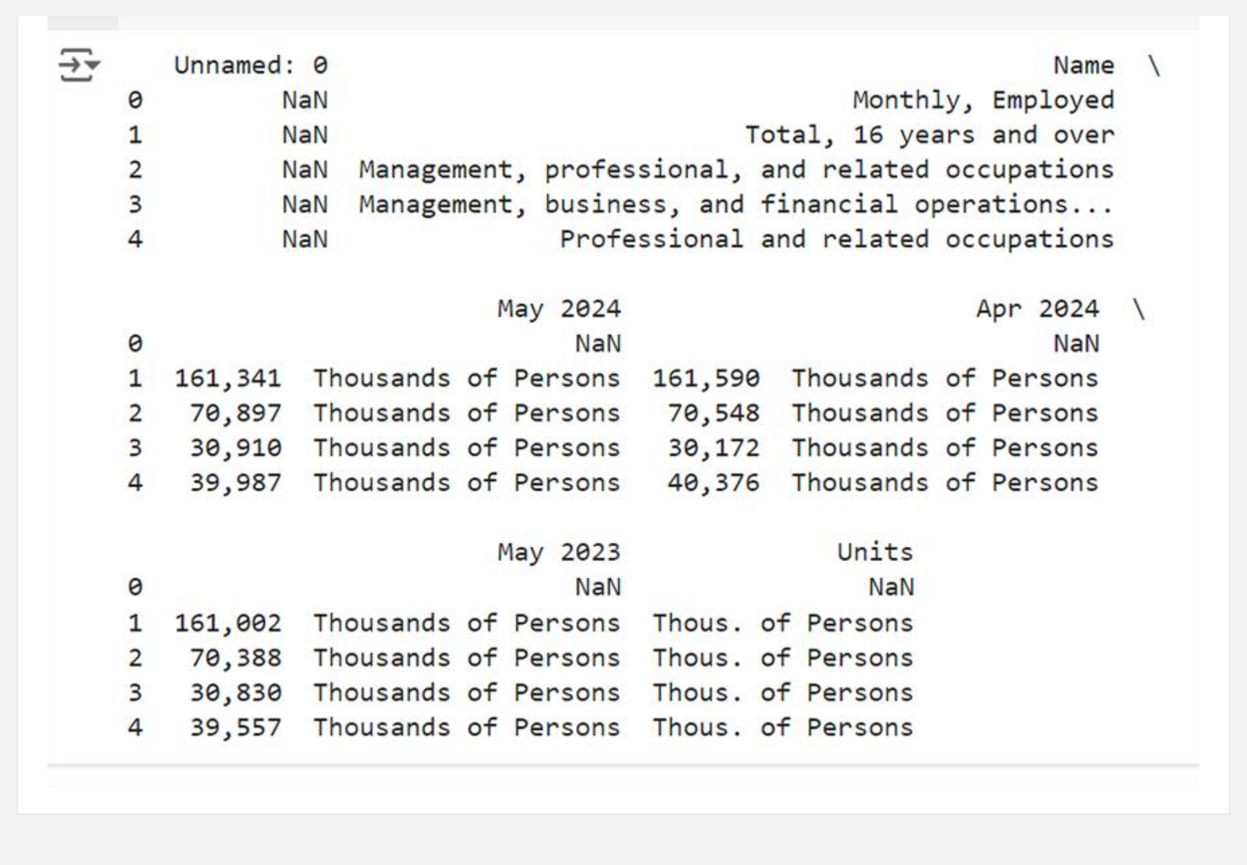
```
# Import pandas
import pandas as p

# Read data from the URL
df_list = pd.read_html(
"https://fred.stlouisfed.org/release/tables?rid=50&eid=3149#snid=4498")

# Since pd.read_html() returns a list of DataFrames, select the first DataFrame
df = df_list[0]

# Print the first 5 rows of the DataFrame
print(df.head(5)
```

The resulting output will look like



```
   Unnamed: 0                                               Name  \
0         NaN                                  Monthly, Employed   
1         NaN                           Total, 16 years and over   
2         NaN  Management, professional, and related occupations   
3         NaN  Management, business, and financial operations...   
4         NaN               Professional and related occupations   

                        May 2024                       Apr 2024  \
0                            NaN                            NaN   
1  161,341  Thousands of Persons  161,590  Thousands of Persons   
2   70,897  Thousands of Persons   70,548  Thousands of Persons   
3   30,910  Thousands of Persons   30,172  Thousands of Persons   
4   39,987  Thousands of Persons   40,376  Thousands of Persons   

                        May 2023              Units  
0                            NaN                NaN  
1  161,002  Thousands of Persons  Thous. of Persons  
2   70,388  Thousands of Persons  Thous. of Persons  
3   30,830  Thousands of Persons  Thous. of Persons  
4   39,557  Thousands of Persons  Thous. of Persons  
```

In Python, there are several libraries and methods that can be used for parsing and extracting data from text. These include the following:

1. Regular expressions (regex or RE) (https://openstax.org/r/docpython). This is a built-in library in Python that allows for pattern matching and extraction of data from strings. It uses a specific syntax to define patterns and rules for data extraction.

2. Beautiful Soup (https://openstax.org/r/pypi). This is an external library that is mostly used for scraping and parsing HTML and XML code. It can be utilized to extract specific data from web pages or documents.

3. Natural Language Toolkit (NLTK) (https://openstax.org/r/nltk). This is a powerful library for natural language processing in Python. It provides various tools for tokenizing, parsing, and extracting data from text data. (Tokenizing is the process of breaking down a piece of text or string of characters into smaller units called tokens, which can be words, phrases, symbols, or individual characters.)

4. TextBlob (https://openstax.org/r/textblob). This library provides a simple interface for most natural language processing assignments, such as argument and part-of-speech tagging. It can also be utilized for parsing and extracting data from text.

5. SpaCy (https://openstax.org/r/spacy). This is a popular open-source library for natural language processing. It provides efficient methods for tokenizing, parsing, and extracting data from text data.

Overall, the library or method used for parsing and extracting data will depend on the specific task and type of data being analyzed. It is important to research and determine the best approach for a given project.

Regular Expressions in Python

Regular expressions, also known as regex, are a set of symbols used to define a search pattern in text data. In Python, these expressions are supported by the re module (function), and their syntax is similar to other programming languages. The use of regular expressions offers researchers a robust method for identifying and manipulating patterns in text. With this powerful tool, specific words, characters, or patterns of characters can be searched and matched. Typical applications include data parsing, input validation, and extracting targeted information from larger text sources. Common use cases in Python involve recognizing various types of data, such as dates, email addresses, phone numbers, and URLs, within extensive text files. Moreover, regular expressions are valuable for tasks like data cleaning and text processing. Despite their versatility, regular expressions can be elaborate, allowing for advanced search patterns utilizing meta-characters like *, ?, and +. However, working with these expressions can present challenges, as they require a thorough understanding and careful debugging to ensure successful implementation.

USING META-CHARACTERS IN REGULAR EXPRESSIONS

- The * character is known as the “star” or “asterisk” and is used to match zero or more occurrences of the preceding character or group in a regular expression. For example, the regular expression "a*" would match an "a" followed by any number (including zero) of additional "a"s, such as "a", "aa", "aaa", etc.

- The ? character is known as the "question mark" and is used to indicate that the preceding character or group is optional. It matches either zero or one occurrences of the preceding character or group. For example, the regular expression "a?b" would match either "ab" or "b".

- The + character is known as the "plus sign" and is used to match one or more occurrences of the preceding character or group. For example, the regular expression "a+b" would match one or more "a"s followed by a "b", such as "ab", "aab", "aaab", etc. If there are no "a"s, the match will fail. This is different from the * character, which would match zero or more "a"s followed by a "b", allowing for a possible match without any "a"s.

EXAMPLE 2.5

Proble

Write Python code using regular expressions to search for a selected word “Python” in a g
print the number of times it

Solutio

PYTHON CODE

```
## import the regular expression module
import r
# create a string with story problem
story ="Samantha is a sixth-grade student who uses the popular coding lang
Python to collect and analyze weather data for science project. She creat
program that collects data from an online weather API and stores it in a
for 10 days. With the help of her teacher, she uses Python to visualize t
and discovers patterns in temperature, humidity, and precipitation."
# create a regex pattern to match a selected word
pattern ="Python"
```

```
# use the "findall" to search for matching patterns in the data string
words = re.findall(pattern, story)

# print the number of repeated python
print("The word 'python' is repeated",len(words),"times in the story.")

# print the matched patterns
print(words)# output: ['Python']
```

The resulting output will look like

```
The word 'python' is repeated 2 times in the story.
['Python', 'Python']
```

In Python, "len" is short for "length" and is used to determine the number of items in a
returns the total count of items in the collection, including spaces and punctuation mar

Parsing and Extracting Data

Splitting and slicing are two methods used to manipulate text strings in programming. Splitting a string means dividing a text string into smaller parts or substrings based on a specified separator. The separator can be a character, string, or regular expression. This can be useful for separating words, phrases, or data values within a larger string. For example, the string "Data Science" can be split into two substrings "Data" and "Science" by using a space as the separator.

Slicing a string refers to extracting a portion or section of a string based on a specified range of indices. An index refers to the position of a character in a string, starting from 0 for the first character. The range specifies the start and end indices for the slice, and the resulting substring includes all characters within that range. For example, the string "Data Science" can be sliced to extract "Data" by specifying the range from index 0 to 4, which includes the first four characters. Slicing can also be used to manipulate strings by replacing, deleting, or inserting new content into specific positions within the string.

Parsing and extracting data involves the analysis of a given dataset or string to extract specific pieces of information. This is accomplished using various techniques and functions, such as splitting and slicing strings, which allow for the structured retrieval of data. This process is particularly valuable when working with large and complex datasets, as it provides a more efficient means of locating desired data compared to traditional search methods. Note that parsing and extracting data differs from the use of regular expressions, as regular expressions serve as a specialized tool for pattern matching and text manipulation. In contrast, parsing and data extraction offers a comprehensive approach to identifying and extracting specific data within a dataset.

Parsing and extracting data using Python involves using the programming language to locate and extract specific information from a given text. This is achieved by utilizing the re library, which enables the use of regular expressions to identify and retrieve data based on defined patterns. This process can be demonstrated through an example of extracting data related to a person purchasing an iPhone at an Apple store.

The code in the following Python feature box uses regular expressions (regex) to match and extract specific data from a string. The string is a paragraph containing information about a person purchasing a new phone from the Apple store. The objective is to extract the product name, model, and price of the phone. First, the code starts by importing the necessary library for using regular expressions. Then, the string data is defined as a variable. Next, regex is used to search for specific patterns in the string. The first pattern searches for the words "product: " and captures anything that comes after it until it reaches a comma. The result is then stored in a variable named "product". Similarly, the second pattern looks for the words "model: " and captures anything that comes after it until it reaches a comma. The result is saved in a variable named "model". Finally, the third pattern searches for the words "price: " and captures any sequence of numbers or symbols that follows it until the end of the string. The result is saved in a variable named "price". After all the data is extracted, it is printed out to the screen, using concatenation to add appropriate labels before each variable.

This application of Python code demonstrates the effective use of regex and the re library to parse and extract specific data from a given data. By using this method, the desired information can be easily located and retrieved for further analysis or use.

PYTHON CODE

```
## import necessary library

import re

# Define data to be parsed data = "Samantha went to the Apple store to purchase a new phone. She was specifically looking for the latest and most expensive model available. As she looked at the different options, she came across the product: iPhone 12, the product name caught her attention, as it was the newest version on the market. She then noticed the model: A2172, which confirmed that this was indeed the latest and most expensive model she was looking for. The price made her hesitate for a moment, but she decided that it was worth it price: $799. She purchased the iPhone 12 and was excited to show off her new phone to her friends."

# Use regex to match and extract data based on specific pattern product = re.search(r"product: (.+?),", data).group(1) model = re.search(r"model: (.+?),", data).group(1) price = re.search(r"price: (.+?.+?.+?.+?)", data).group(1)

# Print the extracted data print("product: " + product) print("model: " + model) print("price: " + price)
```

The resulting output will look like this:

```
product: iPhone 12 model: A2172 price: $799
```

Processing and Storing Data

Once the data is collected, it should be processed and stored in a suitable format for further analysis. This is where data processing and storage come into play. Data processing manages the raw data first by cleaning it through the removal of irrelevant information and then by transforming it into a structured format. The cleaning process includes identifying and correcting any errors, inconsistencies, and missing values in a dataset and is essential for ensuring that the data is accurate, reliable, and usable for analysis or other purposes. Python is often utilized for data processing due to its flexibility and ease of use, and it offers a wide range of tools and libraries specifically designed for data processing. Once the data is processed, it needs to be stored for future use. (We will cover data storage in Data Cleaning and Preprocessing.) Python has several libraries that allow for efficient storage and manipulation of data in the form of DataFrames.

One method of storing data using Python is using the pandas library to create a DataFrame and then using the to_csv() function to save the DataFrame as a CSV (comma-separated values) file. This file can then be easily opened and accessed for future analysis or visualization. For example, the code in the following Python sidebar is a Python script that creates a dictionarywith data about the presidents of the United States (https://openstax.org/r/wikipedia), including their ordered number and state of birth. A data dictionary is a data structure that stores data in key-value pairs, allowing for efficient retrieval of data using its key. It then uses the built-in CSV library (https://openstax.org/r/librarycsv) to create a CSV file and write the data to it. This code is used to store the US presidents' data in a structured format for future use, analysis, or display.

PYTHON CODE

```python
import csv

# Create a dictionary to store the data
presidents = {
    "1": ["George Washington", "Virginia"],
    "2": ["John Adams", "Massachusetts"],
    "3": ["Thomas Jefferson", "Virginia"],
    "4": ["James Madison", "Virginia"],
    "5": ["James Monroe", "Virginia"],
    "6": ["John Quincy Adams", "Massachusetts"],
    "7": ["Andrew Jackson", "South Carolina"],
    "8": ["Martin Van Buren", "New York"],
    "9": ["William Henry Harrison", "Virginia"],
    "10": ["John Tyler", "Virginia"],
    "11": ["James K. Polk", "North Carolina"],
    "12": ["Zachary Taylor", "Virginia"],
    "13": ["Millard Fillmore", "New York"],
    "14": ["Franklin Pierce", "New Hampshire"],
    "15": ["James Buchanan", "Pennsylvania"],
    "16": ["Abraham Lincoln", "Kentucky"],
    "17": ["Andrew Johnson", "North Carolina"],
    "18": ["Ulysses S. Grant", "Ohio"],
    "19": ["Rutherford B. Hayes", "Ohio"],
    "20": ["James A. Garfield", "Ohio"],
    "21": ["Chester A. Arthur", "Vermont"],
    "22": ["Grover Cleveland", "New Jersey"],
    "23": ["Benjamin Harrison", "Ohio"],
    "24": ["Grover Cleveland", "New Jersey"],
    "25": ["William McKinley", "Ohio"],
"26": ["Theodore Roosevelt", "New York"],

"27": ["William Howard Taft", "Ohio"],

"28": ["Woodrow Wilson", "Virginia"], "29": ["Warren G.
Harding", "Ohio"],

"30": ["Calvin Coolidge", "Vermont"],

"31": ["Herbert Hoover", "Iowa"],

"32": ["Franklin D. Roosevelt", "New York"],

"33": ["Harry S. Truman", "Missouri"],

"34": ["Dwight D. Eisenhower", "Texas"],

"35": ["John F. Kennedy", "Massachusetts"],

"36": ["Lyndon B. Johnson", "Texas"],

"37": ["Richard Nixon", "California"],

"38": ["Gerald Ford", "Nebraska"],

"39": ["Jimmy Carter", "Georgia"],

"40": ["Ronald Reagan", "Illinois"],

"41": ["George H. W. Bush", "Massachusetts"],

"42": ["Bill Clinton", "Arkansas"],

"43": ["George W. Bush", "Connecticut"],

"44": ["Barack Obama", "Hawaii"],

"45": ["Donald Trump", "New York"],

"46": ["Joe Biden", "Pennsylvania"] }

# Open a new CSV file in write mode with
open("presidents.csv", "w", newline='') as csv_file:

# Specify the fieldnames for the columns fieldnames =
["Number", "Name", "State of Birth"]

# Create a writer object writer = csv.DictWriter(csv_file,
fieldnames=fieldnames)

# Write the header row writer.writeheader()
# Loop through the presidents dictionary for key, value in
presidents.items():

# Write the data for each president to the CSV file
writer.writerow({

"Number": key,

"Name": value[0],

"State of Birth": value[1]

})

# Print a success message print("Data successfully stored in
CSV file.")
```

The resulting output will look like this:

```
Data successfully stored in CSV file.
```

2.4 Data Cleaning and Preprocessing

Learning Outcomes

By the end of this section, you should be able to:

- 2.4.1 Apply methods to deal with missing data and outliers.
- 2.4.2 Explain data standardization techniques, such as normalization, transformation, and aggregation.
- 2.4.3 Identify sources of noise in data and apply various data preprocessing methods to reduce noise.

Data cleaning and preprocessing is an important stage in any data science task. It refers to the technique of organizing and converting raw data into usable structures for further analysis. It involves extracting irrelevant or duplicate data, handling missing values, and correcting errors or inconsistencies. This ensures that the data is accurate, comprehensive, and ready for analysis. Data cleaning and preprocessing typically involve the following steps:

1. Data integration. Data integration refers to merging data from multiple sources into a single dataset.
2. Data cleaning. In this step, data is assessed for any errors or inconsistencies, and appropriate actions are taken to correct them. This may include removing duplicate values, handling missing data, and correcting formatting misconceptions.
3. Data transformation. This step prepares the data for the next step by transforming the data into a format that is suitable for further analysis. This may involve converting data types, scaling or normalizing numerical data, or encoding categorical variables.
4. Data reduction. If the dataset contains a large number of columns or features, data reduction techniques may be used to select only the most appropriate ones for analysis.
5. Data discretization. Data discretization involves grouping continuous data into categories or ranges, which can help facilitate analysis.
6. Data sampling. In some cases, the data may be too large to analyze in its entirety. In such cases, a sample of the data can be taken for analysis while still maintaining the overall characteristics of the original dataset.

The goal of data cleaning and preprocessing is to guarantee that the data used for analysis is accurate, consistent, and relevant. It helps to improve the quality of the results and increase the efficiency of the analysis process. A well-prepared dataset can lead to more accurate insights and better decision-making.

Handling Missing Data and Outliers

Missing data refers to any data points or values that are not present in a dataset. This could be due to data collection errors, data corruption, or nonresponse from participants in a study. Missing data can impact the accuracy and validity of an analysis, as it reduces the sample size and potentially introduces bias.

Some specific examples of missing data include the following:

1. A survey participant forgetting to answer a question
2. A malfunction in data collection equipment resulting in missing values
3. A participant choosing not to answer a question due to sensitivity or discomfort

An outlier is a data point that differs significantly from other data points in a given dataset. This can be due to human error, measurement error, or a true outlier value in the data. Outliers can skew statistical analysis and bias results, which is why it is important to identify and handle them properly before analysis.

Missing data and outliers are common problems that can affect the accuracy and reliability of results. It is important to identify and handle these issues properly to ensure the integrity of the data and the validity of the analysis. You will find more details about outliers in Measures of Center, but here we summarize the measures typically used to handle missing data and outliers in a data science project:

1. Identify the missing data and outliers. The first stage is to identify which data points are missing or appear to be outliers. This can be done through visualization techniques, such as scatterplots, box plots, IQR (interquartile range), or histograms, or through statistical methods, such as calculating the mean, median, and standard deviation (see Measures of Center and Measures of Variation as well as Encoding Univariate Data).

It is important to distinguish between different types of missing data. MCAR (missing completely at random) data is missing data not related to any other variables, with no underlying cause for its absence. Consider data collection with a survey asking about driving habits. One of the demographic questions asks for income level. Some respondents accidentally skip this question, and so there is missing data for income, but this is not related to the variables being collected related to driving habits.

MAR (missing at random) data is missing data related to other variables but not to any unknown or unmeasured variables. As an example, during data collection, a survey is sent to respondents and the survey asks about loneliness. One of the questions asks about memory retention. Some older respondents might skip this question since they may be unwilling to share this type of information. The likelihood of missing data for loneliness factors is related to age (older respondents). Thus, the missing data is related to an observed variable of age but not directly related to loneliness measurements.

MNAR (missing not at random) data refers to a situation in which the absence of data depends on observed data but not on unobserved data. For example, during data collection, a survey is sent to respondents and the survey asks about debt levels. One of the questions asks about outstanding debt that the customers have such as credit card debt. Some respondents with high credit card debt are less likely to respond to certain questions. Here the missing credit card information is related to the unobserved debt levels.

2. Determine the reasons behind missing data and outliers. It is helpful to understand the reasons behind the missing data and outliers. Some expected reasons for missing data include measurement errors, human error, or data not being collected for a particular variable. Similarly, outliers can be caused by incorrect data entry, measurement errors, or genuine extreme values in the data.

3. Determine how to solve missing data issues. Several approaches can be utilized to handle missing data.

One option is to withdraw the missing data altogether, but this can lead to a loss of important information. Other methods include imputation, where the absent values are replaced with estimated values based on the remaining data or using predictive models to fill in the missing values.

4. Consider the influence of outliers. Outliers can greatly affect the results of the analysis, so it is important to carefully consider their impact. One approach is to delete the outliers from the dataset, but this can also lead to a loss of valuable information. Another option is to deal with the outliers as an individual group and analyze them separately from the rest of the data.

5. Use robust statistical methods. When dealing with outliers, it is important to use statistical methods that are not affected by extreme values. This includes using median instead of mean and using nonparametric tests instead of parametric tests, as explained in Statistical Inference and Confidence Intervals.

6. Validate the results. After handling the missing data and outliers, it is important to validate the results to ensure that they are robust and accurate. This can be done through various methods, such as crossvalidation or comparing the results to external data sources.

Handling missing data and outliers in a data science task requires careful consideration and appropriate methods. It is important to understand the reasons behind these issues and to carefully document the process to ensure the validity of the results.

EXAMPLE 2.6

Proble

Starting in 1939, the United States Bureau of Labor Statistics tracked employment on a m
number of employers in the construction field between 1939 and 2019 is Figure .

a Determine if there is any outlier in the dataset that deviates significantly from the overa
b In the event that the outlier is not a reflection of real employment numbers, how would
outliers

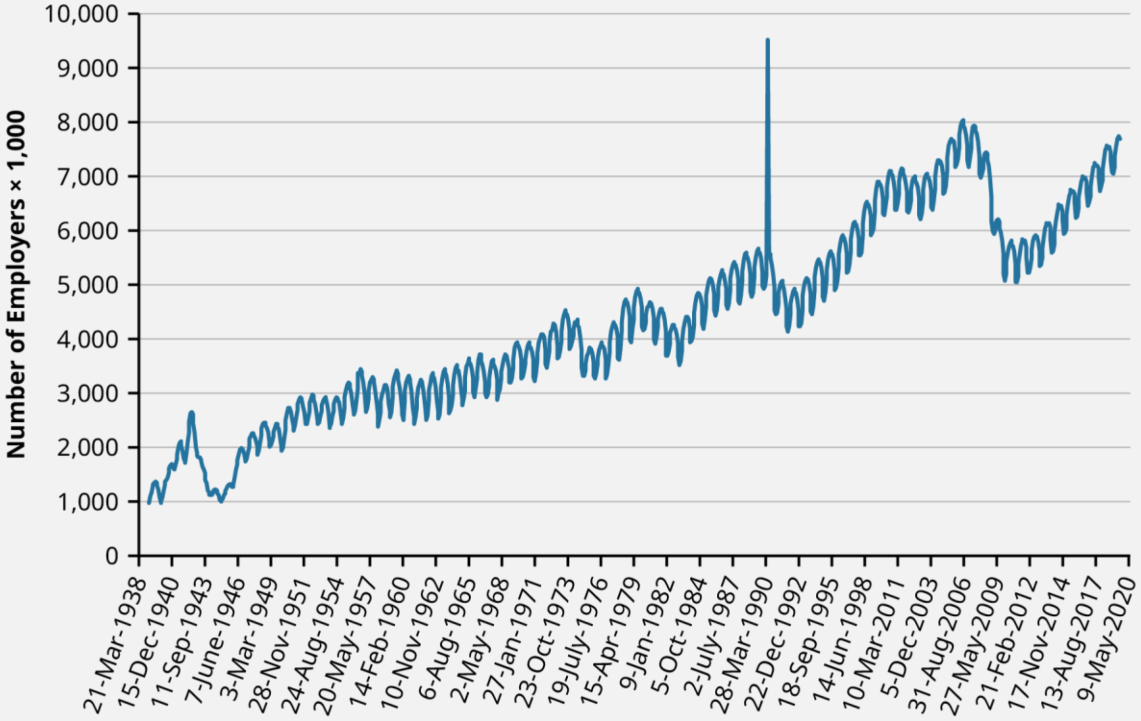


Figure 2.2 US Retail Employment with Outliers
(Data source: Bureau of Labor Statistics; credit: modification of work by Hyndman, R.J., & Athanasopoulos, G. (2021). Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on April 23, 2024)

Solutio

a Based on the visual evidence, it appears that an outlier is present in the employment d
199. The employment level during this month shows a significant jump from approxima
9,500, exceeding the overall maximum employment level recorded between 1939 2019
b A possible approach to addressing this outlier is to replace it with median, calculated
mean of the points before and after the outlier. This method can help to improve the
realism of the employment curve as well as mitigate any potential impacts the outlier r
statistical analysis or modeling processes. Table .

Dat	Number Employers ×
1/1/199	497
2/1/199	493
3/1/199	498

Dat	Number Employers ×
4/1/199	517
5/1/199	537
6/1/199	952
7/1/199	558
8/1/199	558
9/1/199	548
1 /1/199	537
1 /1/199	520
1 /1/199	496

Table 2.3 Monthly Employment over 1990

The median employment level from May 1, 1990, to July 1, 1990, is 5,289, as a replacemen outlier on May 28,

The adjusted data is presented Figure.

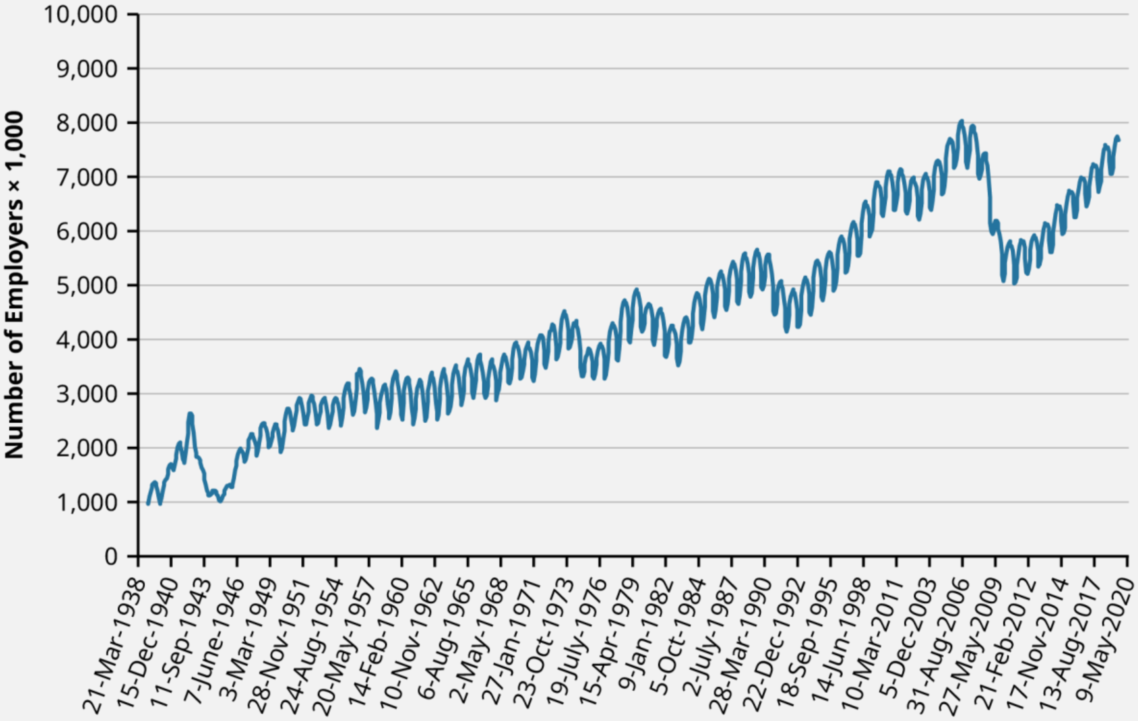


Figure 2.3 US Retail Employment Without Outliers
(Data source: Bureau of Labor Statistics; credit: modification of work by Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on April 23, 2024)

Data Standardization, Transformation, and Validation

Data standardization, transformation, and validation are critical steps in the data analysis preprocessing pipeline. Data standardization is the process of systematically transforming collected information into a consistent and manageable format. This procedure involves the elimination of inconsistencies, errors, and duplicates as well as converting data from various sources into a unified format, often termed a normal form (defined in the next section). Data transformation involves modifying the data to make it more suitable for the analysis that is planned. Data validation ensures that the data is accurate and consistent and meets certain criteria or standards. Table 2.4 summarizes these processes, which are explored in more depth later in the chapter.

Purpose Techniques Example

Normalization To eliminate Involve breaking A customer database table with
redundant data down a larger redundant data can be

and ensure data database into smaller, normalized by splitting it into
dependencies more manageable two tables for customer make sense tables
and information and orders and

establishing establishing the relationship relationships between them. between them.

Transformation To make the data Include data A dataset with different date

more consistent, cleaning, data formats, such as “MM/DD/YYYY” coherent, and
merging, data and “YYYY/MM/DD.” Data meaningful for splitting, data
transformation can be used to

analysis. conversion, and data convert all dates to a single aggregation. format

Validation To improve the Include data In a survey, the respondent’s age

quality of data profiling, data audits, is recorded as 150 years. used in analysis.
It and data cleansing. Through data validation, this

ensures that the value can be identified as data is accurate, erroneous. relevant,
and consistent.

Table 2.4 Comparison of Data Standardization, Transformation, and Validation Processes

Data Normalization

The first step in standardizing data is to establish guidelines and rules for formatting and structuring the data. This may include setting conventions for naming, data types, and formatting A normal form (NF) is a guideline or set of rules used in database design to ensure that a database is well-structured, organized, and free from certain types of data irregularities, such as redundancy and inconsistency. The most commonly used normal forms are 1NF, 2NF, 3NF(First, Second, and Third Normal Form), and BCNF(Boyce-Codd Normal Form).

Normalization is the process of applying these rules to a database. The data must be organized and cleaned, which involves removing duplicates and erroneous data, filling in missing values, and logically arranging the data. To uphold data standardization, regular quality control measures should be implemented, including periodic data audits to ascertain the accuracy and consistency of the data. It is also important to document the standardization process, including the guidelines and procedures followed. Periodic review and updates of data standards are necessary to ensure the ongoing reliability and relevance of the data.

Data normalization ensures that data is maintainable regardless of its source. Consider a marketing team collecting data on their customers' purchasing behavior so they can make some decisions about product placement. The data is collected from numerous sources, such as online sales transactions, in-store purchases, and customer feedback surveys. In its raw form, this data could be disorganized and unreliable, making it difficult to analyze. It is hard to draw meaningful insights from badly organized data.

To normalize this data, the marketing team would go through multiple steps. First, they would identify the key data elements, such as customer name, product purchased, and transaction date. Then, they would ensure that these elements are consistently formatted in all data sources. For instance, they might employ the same date format across all data sources or standardize customer names to the first name and the last name fields. Subsequently, they would eliminate any redundant or irrelevant data elements. In this case, if the data is collected from both online and in-store purchases, they might choose one or the other to avoid duplication. The marketing team would ensure that the data is properly structured and organized. This could involve creating a data table with domains for each data element, such as a customer ID, product code, and purchase amount. By normalizing the data, the marketing team can efficiently follow and investigate the customers' purchasing behavior, identify patterns and trends, and make data-driven judgments to enhance their marketing systems.

The normalization formula is a statistical formula utilized to scale down a dataset, typically to between one and zero. The largest data would have a normalized value of one, and the smallest data point would be zero. Note that the presence of outliers can significantly impact the calculated values of minimum/maximum. As such, it is important to first remove any outliers from the dataset before performing normalization. This ensures more accurate and representative results.

The normalization

$$x_{\text{norm}} = \frac{x - \min}{\max - \min}$$

EXAMPLE 2.7

Proble

A retail company with eight branches across the country wants to analyze its product sale
top-selling items. The company collects data from each Table branch, and listing the
and profits for each product. From previous reports, it discovered that its top-selling produ
TV accessories, beauty products, DVDs, kids' toys, video games, women's boutique appar
and fashion sunglasses. However, the company wants to arrange these products in order
lowest based on best sales and profits. Determine which product is the top-selling produ
the data Table .

Branc	Produc	Sales	Profits
Branch	Jewelr	5000	2000
Branch	TV	2500	1250
Branch	Beauty	3000	1500
Branch	DVD	1500	750
Branch	Kids’	4500	2250
Branch	Video	3500	1750
Branch	Women Boutiqu Appare	4000	2000
Branch	Designer & Sunglass	5500	2750

Table 2.5 Retail Company Sales and Profits

Solutio

Using the normalization formula, the maximum sale is $55,000, and the minimum sale is
i Table .

Branc	Produc	Sales	Profits	Normalizatio Scal
Branc1	Jewelr	5000	2000	0.8
Branc2	TV	2500	1250	0.2
Branc3	Beauty	3000	1500	0.3
Branc4	DVD	1500	750	0.0
Branc5	Kids’	4500	2250	0.7
Branc6	Video	3500	1750	0.5
Branc7	WomenBoutiquAppare	4000	2000	0.6
Branc8	Designer & Sunglass	5500	2750	1.0

Table 2.6 Data of Retail Company Sales on Normalization Scale

Overall, the retail company’s top-selling products generate the most profits for the compa
& Fashion Sunglasses” being the highest in the normalization scale. The company can us
to focus on promoting and restocking these items at each branch to continue driving sale

Data Transformation

Data transformation is a statistical technique used to modify the original structure of the data in order to make it more suitable for analysis. Data transformation can involve various mathematical operations such as logarithmic, square root, or exponential transformations. One of the main reasons for transforming data is to address issues related to statistical assumptions. For example, some statistical models assume that the data is normally distributed. If the data is not distributed normally, this can lead to incorrect results and interpretations. In such cases, transforming the data can help to make it closer to a normal distribution and improve the accuracy of the analysis.

One commonly used data transformation technique is the log transformation, which requires taking the logarithm of the data values. Log transformation is often used when the data is highly skewed, meaning most of the data points fall toward one end of the distribution. This can cause problems in data analysis as the data may not follow a normal distribution. By taking the logarithm of the values, the distribution can be shifted toward a more symmetrical shape, making it easier to analyze. Another common transformation technique is the square root transformation, which involves taking the square root of the data values. Similar to log transformation, square root transformation is often used to address issues with skewness and make the data more normally distributed. Square root transformation is also useful when the data contains values close to zero, as taking the square root of these values can bring them closer to the rest of the data and reduce the impact of extreme values. Exponential transformations involve taking the exponent of the data values. Whichever operation is used, data transformation can be a useful tool for data analysts to address issues with data distribution and improve the accuracy of their analyses.

Dealing with Noisy Data

Noisy data refers to data that retains errors, outliers, or irrelevant information that can conceal true patterns and relationships within the dataset. The presence of noisy data in the dataset causes difficulty in drawing accurate conclusions and making predictions from the data. Most noisy data is caused by human errors in data entry, technical errors in data collection or transmission, or natural variability in the data itself. Noisy data is removed and cleaned by identifying and correcting errors, removing outliers, and filtering out irrelevant information. Noisy data can negatively impact data analysis and modeling, and it may indicate that there are issues with the model's structure or assumptions. Noisy data is unwanted information that can be removed.

Strategies to reduce the noisy data are summarized in Table 2.7.

Strategy	Example
Data cleaning	Removing duplicate or irrelevant data from a dataset, such as deleting repeated rows in a spreadsheet or filtering out incomplete or error-prone data entries.
Data smoothing	A technique used in data analysis to remove outliers or noise from a dataset in order to reveal underlying patterns or trends. One example is smoothing a dataset of daily stock market index values over the course of a month. The values may fluctuate greatly on a day-to-day basis, making it difficult to see any overall trend. By calculating a seven-day average, we can smooth out these fluctuations and see the overall trend of the index over the course of the month.
Imputation	An example of imputation is in a hospital setting where a patient's medical history is incomplete due to missing information. The hospital staff can use imputation to estimate the missing data based on the patient's known medical conditions and past treatments.
Binning	A researcher is studying the age demographics of a population in a city. Instead of looking at individual ages, the researcher decides to bin the data into age groups of 10 years (e.g., 0–10, 10–20, 20–30, etc.). This allows for a more comprehensive and easier analysis of the data.
Data transformation	Consider the following dataset that shows the number of COVID-19 cases recorded in a country at different points in time—01/01/2020, 02/02/2020, 03/

	01/2020—are 1000, 10000, 100000, respectively. To transform this data using log, we can take the log base 10 of the number of cases column. This would result in a transformed data as follows: 01/01/2020, 02/02/2020, 03/01/2020 are 3, 4, 5, respectively.
Dimensionality reduction	The original dataset would have high dimensionality due to the large number of variables (100 companies) and time points (5 years). By applying principal component analysis, we can reduce the dimensionality of the dataset to a few principal components that represent the overall trends and patterns in the stock market.
Ensemble methods	An example of an ensemble method is the random forest algorithm. It combines multiple decision trees, each trained on a random subset of the data, to make a more accurate prediction. This helps reduce overfitting and increase the overall performance of the model. The final prediction is made by aggregating the predictions of each individual tree.

Table 2.7 Strategies to Reduce Noisy Data

Data Validation

Data validation is the process of ensuring the accuracy and quality of data examined against defined rules and standards. This approach involves identifying and correcting any errors or inconsistencies in the collected data as well as ensuring that the data is relevant and reliable for analysis. Data validation can be performed through a variety of techniques, including manual checks, automated scripts, and statistical analysis. Some common inspections in data validation include checking for duplicates, checking for mislaid values, and verifying data against external sources or references. Before collecting data, it is important to determine the conditions or criteria that the data needs to meet to be considered valid. This can include factors such as precision, completeness, consistency, and timeliness. For example, a company may set up a data validation process to ensure that all customer information entered into its database follows a specific format. This would involve checking for correct spellings and proper formatting of phone numbers and addresses and validating the correctness of customer names and account numbers. The data would also be checked against external sources, such as official government records, to verify the accuracy of the information. Any discrepancies or errors would be flagged for correction before the data is used for analysis or decision-making purposes. Through this data validation process, the company can ensure that its customer data is accurate, reliable, and compliant with industry standards.

Another method to assess the data is to cross-reference it with reliable sources to identify any discrepancies or errors in the collected data. A number of tools and techniques can be used to validate data. These can include statistical analysis, data sampling, data profiling, and data auditing. It is important to identify and remove outliers before validating the data. Reasonability checks involve using common sense to check if the data is logical and makes sense—for example, checking if a person's age is within a reasonable range or if a company's revenue is within a reasonable range for its industry. If possible, data should be verified with the source to ensure its accuracy. This can involve contacting the person or organization who provided the data or checking against official records. It is always a good idea to involve multiple team members or experts in the validation process to catch any errors or inconsistencies that may have been overlooked by a single person. Documentation of the validation process, including the steps taken and any issues identified, is important in future data audits or for reference purposes. Data validation is a continuous process, and data should be monitored and updated to ensure its accuracy and validity.

Consider a marketing company conducting a survey on customer satisfaction for a new product launch. The company collected data from 1,000 respondents, but when the company started analyzing the data, it noticed several inconsistencies and missing values. The company's data analyst realized that the data standardization and validation processes were not adequately performed before the survey results were recorded. To correct this issue, the data analyst first identified and removed all duplicate entries, reducing the total number of responses to 900. Then, they used automated scripts to identify and fill in missing values, which accounted for 95 responses. The remaining 805 responses were then checked for data accuracy using statistical analysis. After the data standardization and validation process, the company had a clean and reliable dataset of 805 responses. The results showed that the product had a satisfaction rate of 85%, which was significantly higher than the initial analysis of 78%. As a result of this correction, the marketing team was able to confidently report the actual satisfaction rate and make better-informed decisions for future product development.

Data Aggregation

Data aggregation is the process by which information from multiple origins is gathered and merged into a single set that provides insights and meaningful conclusions. It involves gathering, managing, and delivering data from different sources in a structured manner to facilitate analysis and decision-making. Data aggregation can be performed manually or by using automated tools and techniques. The data aggregation process is utilized to identify patterns and trends between different data points, which extracts valuable insights. Some standard types of data aggregation are spatial aggregation, statistical aggregation, attribute aggregation, and temporal aggregation. This methodology is commonly operated in marketing, finance, health care, and research to analyze large sets of data. Data aggregation is used in various industries to combine and analyze large sets of data. Examples include calculating total sales for a company from different departments, determining average temperature for a region including multiple cities, and analyzing website traffic by country. It is also used in fields such as stock market indices, population growth, customer satisfaction scores, credit scores, and airline flight delays. Governments and utility companies also utilize data aggregation to study energy consumption patterns.

Text Preprocessing

Text preprocessing is a technique of preparing data in text format for further analysis and natural language processing tasks. It involves transforming unstructured text data into a more structured format to be interpreted by algorithms and models. Some common techniques used in text preprocessing are reviewed in Table 2.8.

Preprocessing Explanation Example

Technique

Tokenization Breaking text Original Text: “Tokenization is the process of breaking down data into a sentence, paragraph, or entire text into smaller parts

individual called tokens.” words or phrases Tokenized Text: “Tokenization”, “is”, “the”, “process”, “of”, (tokens) “breaking”, “down”, “a”, “sentence”, “,”, “paragraph”, “,”, “or”,

“entire”, “text”, “into”, “smaller”, “parts”, “called”, “tokens”, “.”

Lowercasing Converting all Consider the following sentence: “John likes to eat pizza.” text to After lowercasing it, the sentence becomes “john likes to eat lowercase to pizza.”

avoid multiple representations

of the identical

word

Stopwords removal Filtering out Consider: “The sun was shining bright in the sky and the commonly birds were chirping. It was a lovely day in the park and occurring people were enjoying the beautiful weather.” words that do not add After removing stopwords, the paragraph would be meaning or transformed into: “Sun shining bright sky birds chirping.

context Lovely day park people enjoying beautiful weather.” Lemmatization and Reducing For example, the words “running,” “runs,” and “ran” would all

stemming words to their be lemmatized to the base form ”run.”

root forms to reduce complexity and improve model performance

Table 2.8 Summary of Text Preprocessing Techniques

Preprocessing Explanation Example

Technique

Part-of-speech tagging Identifying the In the sentence “I went to the market to buy some fruits,” grammatical the words “went” and “buy” would be tagged as verbs, components of “market” and “fruits” as a noun, and “some” as adjectives. each word where each word in a sentence is assigned a specific part-ofspeech tag (e.g., noun, verb, or adjective)

Named entity Recognizing Consider the following text: “John went to Paris last summer recognition and with his colleagues at Microsoft.”

categorizing named entities By using named entity recognition, we can tag the named such as people, entities in this sentence as follows: “John (person) went to places, and Paris (location) last summer with his colleagues at Microsoft organizations (organization).”

Sentiment analysis Identifying and Let’s say a customer has left a review for a new laptop they categorizing purchased. The review reads: “I am extremely satisfied with the emotions my purchase. The laptop has exceeded all of my expressed in expectations and has greatly improved my work efficiency. textThanks for an amazing product!”

To perform sentiment analysis, the text will first undergo preprocessing, which involves cleaning and preparing the text data for analysis. This may include removing punctuation, converting all letters to lowercase, and removing stopwords (commonly used words that do not add much meaning to a sentence, such as “the” or “and”).

After preprocessing, sentiment analysis techniques will be applied to analyze the emotions and opinions expressed in the review. The analysis may identify key positive words such as “satisfied” and “amazing” and measure their overall sentiment score. It may also take into account the context and tone of the review to accurately determine the sentiment.

Table 2.8 Summary of Text Preprocessing Techniques

Preprocessing Explanation Example

Technique

Spell-checking and Correcting Suppose we have a text like: “The writting was very correction spelling errors inpprecise and had many trgical mistakes.” With spellto improve checking and correction, this text can be processed and accuracy

corrected to: “The writing was very imprecise and had many tragic mistakes.” This involves identifying and correcting misspelled words.

In this case, “writting” was corrected to “writing,”

“inpprecise” to “imprecise,” and “trgical” to “tragic.” This not only improves the readability and accuracy of the text but also helps in better understanding and analysis of the text.

Encoding Converting text Let’s say we have a dataset of customer reviews for a into a restaurant. Each review is a string of text, such as: “I had a numerical great experience at this restaurant, the food was delicious representation and the service was top-notch.” To encode this text data, we

for machine can use techniques such as one-hot encoding or word learning embedding.

algorithms to

process One-hot encoding involves creating a binary vector for each

word in the review, where the size of the vector is equal to the size of the vocabulary. For example, if the review contains the words “great,” “experience,” “restaurant,” “food,” “delicious,” “service,” and “top-notch,” the one-hot encoded vectors for these words would be: great: [1, 0, 0, 0, 0, 0, 0]; experience: [0, 1, 0, 0, 0, 0, 0]; restaurant: [0, 0, 1, 0, 0, 0, 0]; food: [0, 0, 0, 1, 0, 0, 0]; delicious: [0, 0, 0, 0, 1, 0, 0]; service: [0, 0, 0, 0, 0, 1, 0]; top-notch: [0, 0, 0, 0, 0, 0, 1]. These onehot encoded vectors can now be used as input features for machine learning models.

Removing special Simplifying the Consider the input: “Hello, this is a text with @special#

characters and text for analysis characters!*” and the output: “Hello this is a text with special punctuation characters”

Table 2.8 Summary of Text Preprocessing Techniques

Text preprocessing is crucial to effectively use text data for tasks such as text classification and information extraction. By transforming the raw text data, the accuracy and performance of machine learning models can greatly improve and provide meaningful insights and predictions.

Text preprocessing is especially important in artificial intelligence, as it can lay the foundation for effective text analysis, classification, information retrieval, sentiment analysis, and many other natural language processing tasks (see Natural Language Processing). A well-designed preprocessing pipeline can lead to better model performance, improved efficiency, and more accurate insights from text data.

An example of text preprocessing in artificial intelligence involves the conversion of text data into a more standardized and consistent format. This can include tasks such as removing accents and diacritics, expanding contractions, and converting numbers into their written representations (e.g., “10” to “ten”). Normalizing text data helps to reduce the number of variations and therefore improves the efficiency and accuracy of natural language processing tasks. It also makes the data more easily interpretable for machine learning algorithms. For example, when analyzing customer reviews, it may be beneficial to normalize text data so that variations of the same word (e.g., “colour” and “color”) are treated as the same, providing a more accurate understanding of sentiment toward a product or service.

2.5 Handling Large Datasets

Learning Outcomes

By the end of this section, you should be able to:

- 2.5.1 Recognize the challenges associated with large data, including storage, processing, and analysis limitations.
- 2.5.2 Implement techniques for efficient storage and retrieval of large datasets, including compression, indexing, and chunking.
- 2.5.3 Discuss database management systems and cloud computing and their key characteristics with regard to large datasets.

Large datasets, also known as big data, are extremely large and complex sets of data that traditional data processing methods and tools are unable to handle. These datasets typically include sizeable volumes, variety, and velocity of data, making it challenging to process, manage, and analyze them using traditional methods. Large datasets can be generated by a variety of sources, including social media, sensors, and financial transactions. They generally possess a high degree of complexity and may have structured, unstructured, or semi-structured data. Large datasets are covered in more depth in Other Machine Learning Techniques.

We have also already discussed a number of the techniques and strategies used to gain meaningful insights from big data. Survey Design and Implementation discussed sampling techniques that allow those working with data to analyze and examine large datasets by select a representative subset, or representative random

sample. Preprocessing techniques covered in Data Cleaning and Preprocessing are also used to clean, normalize, and transform data to make sure it is consistent before it can be analyzed. This includes handling missing values, removing outliers, and standardizing data formats.

In this section we will consider several other aspects of data management that are especially useful with big data, including data compression, data storage, data indexing, and chunking. In addition, we’ll discuss database management systems—software that allows for the organization, manipulation, and retrieval of data that is stored in a structured format—and cloud computing.

Data Compression

Data compression is a method of reducing file size while retaining essential information; it can be applied to many types of databases. Data compression is classified into two categories, lossy and lossless:

- Lossy compression reduces the size of data by permanently extracting particular data that is considered irrelevant or redundant. This method can significantly decrease file sizes, but it also results in a loss of some data. Lossy compression is often utilized for multimedia files and images, where slight modifications in quality may not be noticeable to the human eye. Examples of lossy compression include MP3 and JPEG.

- Lossless compression aims to reduce file size without removing any data. This method achieves compression by finding patterns and redundancies in the data and representing them more efficiently. This allows for the reconstruction of the original data without any loss in quality. Lossless compression is commonly used for text and numerical data, where every piece of information is crucial. Examples of lossless compression include ZIP, RAR, and PNG.

There are several methods of data lossless compression, including Huffman coding. Huffman coding works by assigning shorter binary codes to the most frequently used characters or symbols in a given dataset and longer codes to less frequently used characters, as shown in Figure 2.4. This results in a more efficient use of binary digits and reduces the overall size of the data without losing any information during compression. Huffman coding is applied to data that require fast and efficient compression, such as video compression, image compression, and data transmission and storage.

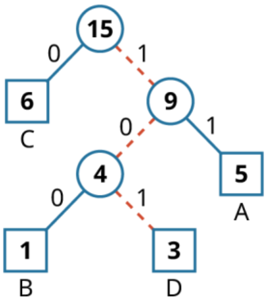


Figure 2.4 Huffman Tree Diagrams

Data Storage

Big data requires storage solutions that can handle large volumes of diverse data types, offer high performance for data access and processing, guarantee scalability for growing datasets, and are thoroughly reliable. The choice of storage solution will depend on data volume, variety, velocity, performance requirements, scalability needs, budget constraints, and existing infrastructure. Organizations often deploy a combination of storage technologies to address different uses and requirements within their big data environments. Some common types of storage solutions used for big data include:

1. Relational databases: Relational databases. organize data in tables and uses structured query language (SQL) for data retrieval and management. They are commonly used for traditional, structured data such as financial data.

2. NoSQL databases. These databases are designed to handle unstructured data, such as social media content or data from sensors, and use non-relational data models.

3. Data warehouses. A data warehouse is a centralized repository of data that combines data from multiple sources and allows for complex queries and analysis. It is commonly used for business intelligence and reporting intentions.

4. Cloud storage. Cloud storage involves storing data in remote servers accessed over the internet. It offers scalability, cost-effectiveness, and remote accessibility.

5. Object storage. With object storage, data are stored as objects that consist of both data and metadata. This method is often used for storing large volumes of unstructured data, such as images and videos.

Data Indexing

Data indexing is an important aspect of data management and makes it easier to retrieve specific data quickly and efficiently. It is a crucial strategy for optimizing the performance of databases and other data storage systems. Data indexing refers to the process of managing and saving the collected and generated data in a database or other data storage system in a way that allows for efficient and fast return of specific data.

Indexing techniques vary in how they organize and store data, but they all aim to improve data retrieval performance. B-tree indexingis illustrated in Figure 2.5. It involves organizing data in a tree-like structure with a root node and branches that contain pointers to other nodes. Each node contains a range of data values and pointers to child nodes, allowing for efficient searching within a specific range of values.

Hashes indexinginvolves using a hash function to map data to a specific index in a table. This allows for direct access to the data based on its hashed value, making retrieval faster than traditional sequential searching. Bitmap indexingis a technique that involves creating a bitmap for each different value in a dataset. The bitmaps are then combined to quickly identify records that match a specific set of values, allowing efficient data retrieval.

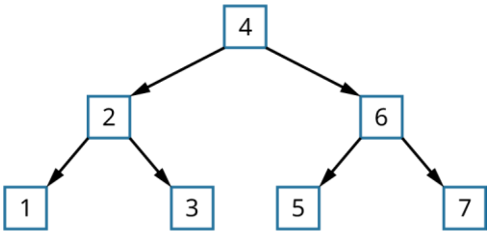


Figure 2.5 B-Tree Indexing

Data Chunking

Data chunking, also known as data segmentationor data partitioning, is a technique used to break down large datasets into smaller, more manageable chunks and make them easier to manage, process, analyze, and store. Chunking is particularly useful when datasets are too large to be processed or analyzed as a single unit. By dividing the data into smaller chunks, various processing tasks can be distributed across multiple computing nodes or processing units.

Data chunking is used in data storage systems and data transmission over networks, and it is especially useful when working with large datasets that exceed the capacity of a single machine or when transferring data over a network with limited bandwidth. The divided data in data chunking is known as a chunkor block. The size of each chunk can vary depending on the requirements and range from a few kilobytes to several gigabytes. These chunks are typically organized sequentially, with each chunk containing a set of data from the larger dataset. The process of data chunking also involves adding metadata (data that provides information about other data), such as the chunk number and the total number of chunks, to each chunk. This metadata allows the chunks to be reassembled into the original dataset after being transmitted or stored separately. Data chunking has several advantages, including the following:

1. Increased speed. By dividing a large dataset into smaller chunks, data processing and transmission can be performed more quickly, reducing the overall processing time.

2. Better utilization of resources. Data chunking enables data to be distributed and processed across multiple machines, making better use of available computing resources.

3. Increased fault tolerance. In case of data corruption or loss, data chunking allows for the retrieval of only the affected chunk rather than the entire dataset.

4. Flexibility. Data chunking allows for the transfer and processing of only the required chunks rather than the entire dataset, providing flexibility in managing large datasets.

Database Management Systems

Database management is a crucial aspect of data science projects as it involves organizing, storing, retrieving, and managing large volumes of data. In a data science project, a database management system (DBMS) is used to ensure the efficient storage and retrieval of data. Database management systems are software tools used for managing data in a structured format. Their functions are summarized in Table 2.9.

	Descriptio	Benefi
Data	Provides a warehouse for storing types of data in a forma	Makes data easy to retrieve and
Data	Allows for efficient and retrieval of data from database using queries filter	Makes data more accessible to scientist

Table 2.9 Functions of Database Management Systems (DBMS)

DBMS Function DBMS Function Description Benefit

Data organization Helps to manage data in a Makes data more manageable for

structured format performing analysis and identifying patterns

or relationships between different data points

Data security	Provides strong security measures to protect sensitive data from unauthorized access	Protects sensitive data such as personal information or financial data from unauthorized access
Data integration	Permits the integration of data from multiple sources	Makes it possible to combine and analyze data from different datasets

Table 2.9 Functions of Database Management Systems (DBMS)

Implementation of database management techniques has become important for hospitals in achieving better patient outcomes and reducing costs. This strategy involves the collection and analysis of patient data from different sources, including electronic health records, medical imaging, and lab results. For example, hospitals are utilizing this approach to improve treatment for patients with chronic conditions such as diabetes and heart disease. By leveraging data-driven insights and identifying patterns, health care experts can develop personalized remedy plans for each patient, guiding them to enhanced health and wellness as well as cost savings for the hospital. This showcases the significance of incorporating advanced data management techniques in health care systems. Through accurate and efficient management and analysis of patient data, hospitals and health care providers are able to make informed decisions, eventually resulting in a more efficient and effective health care system overall.

Cloud Computing

Cloud computing delivers a cost-effective solution for storing vast amounts of data, enabling seamless collaboration and data transfer among remote groups. This technology comprises remote-access tools for storage, processing, and analytics, facilitating multiple users' access regardless of their physical location. Moreover, cloud computing boasts a diverse range of data collection tools, including machine learning, and data warehouses, streamlining the data assembly operation and improving overall efficiency. Cloud computing equips data scientists with the necessary resources and flexibility to effectively collect, manage, and analyze data for their projects. Some examples of cloud storage are Amazon AWS, Microsoft Azure, and Google Cloud.

EXAMPLE 2.8

Problem

The CEO of a large insurance company faced the challenge of increasing digital processes and documents, leading to a need for more storage capacity and rising costs for maintaining servers and hardware. What are all the available options for the insurance company facing the need for additional storage capacity at a higher cost, and which solution would be the most effective for decreasing costs and increasing capacity while ensuring data security?

Solution

1. Migrating to a cloud storage system: This would allow the company to offload the burden of maintaining physical servers and hardware while also providing virtually unlimited storage capacity. The cost of cloud storage is also generally more flexible and scalable, allowing the company to only pay for the storage it needs. Additionally, with the use of cloud-based document management systems, the company can streamline and automate its processes, reducing the need for physical documentation and increasing efficiency. However, this decision would require careful consideration of security measures and potentially training for employees to adapt to the new system.

2. Implementing a data archiving strategy: Instead of immediately migrating to the cloud or investing in new technology, the company could consider implementing a data archiving strategy. This involves identifying and storing infrequently used data in a separate, low-cost storage system, freeing up space on servers and reducing costs.

3. Outsourcing data storage and management: The company could consider outsourcing its data storage and management to a third-party provider. This would eliminate the need for maintaining physical servers and hardware, and the provider may also offer advanced security measures and data backup options.

4. Consolidating data and processes: The company could assess its current processes and systems to identify areas where data and processes can be consolidated to reduce the need for multiple storage systems and streamline workflows.

5. Implementing a virtual desktop infrastructure: A virtual desktop infrastructure allows employees to access their desktop and applications from a central server, reducing the need for individual storage space on devices. This can also improve security and data backup capabilities.

6. Upgrading or redesigning its current storage system: The company could invest in upgrading or redesigning its current storage system to improve efficiency and increase storage capacity.

7. Utilizing hybrid cloud storage: Instead of fully migrating to the cloud, the company could consider a hybrid approach where certain sensitive data is stored on-premises while less critical data is stored in the cloud. This approach can offer the benefits of both on-premises and cloud storage.

8. Conducting regular audits of data usage and storage: The company could conduct regular audits of its data usage and storage to identify areas of redundancy or

inefficiency and adjust accordingly. This can help optimize storage and reduce costs over time.

EXAMPLE 2.9

Problem

What is the descending order of storage capacity in Example 2.8, starting from highest to lowest, while maintaining the same cost and the same level of security?

Solution

The available storage options in Example 2.8, ranked from highest to lowest capacity at the same cost and security level, are:

1. Cloud storage system: This option would provide virtually unlimited storage capacity, as the company can scale up as needed at a relatively low cost. However, the overall capacity and cost would depend on the specific cloud storage plan chosen.

2. Data archiving strategy: By identifying and storing infrequently used data, the company can free up space on its current servers and reduce costs. However, the storage capacity will be limited to the existing servers.

3. Outsourcing data storage and management: By outsourcing to a third-party provider, the company can potentially access higher storage capacity than its current setup, as the provider may have more advanced storage options. However, this would also depend on the specific plan and cost negotiated.

4. Implementing a virtual desktop infrastructure: This option may provide slightly lower storage capacity compared to outsourcing, as it still relies on a central server. However, it can still be an improvement compared to the company's current setup.

5. Upgrading or redesigning its current storage system: This option may also result in lower storage capacity compared to outsourcing, as it only involves improving the existing server setup rather than using an external provider. However, it can still provide a significant increase in capacity depending on the specific upgrades implemented.

6. Hybrid cloud storage: This option can provide a mix of higher and lower storage capacity depending on the specific data stored in the cloud and on-premises. Sensitive data may have lower capacity on the cloud, while less critical data can be stored at higher capacity.

7. Consolidating data and processes: By streamlining processes and eliminating redundancies, the company can potentially reduce the need for excessive storage capacity. However, this would also depend on the company's current setup and level of optimization.

8. Regular audits of data usage and storage: This option may not directly impact storage capacity, but by identifying and eliminating redundancies, the company can optimize its existing storage capacity and potentially reduce costs in the long run.

Datasets

Note: The primary datasets referenced in the chapter code may also bedownloaded here (https://openstax.org/r/drive).)