



데이터

1

디지털 데이터의 압축과 암호화

데이터를 효율적으로 관리하고
보호하는 방법은 무엇일까?



2

빅데이터

빅데이터를 목적에 맞게
처리하여 분석하는
방법은 무엇일까?



“

데이터 압축과 암호화는 데이터를 안전하게 관리하고
보호하는 데 효율적이고, 빅데이터 기술을
활용하면 데이터를 관리하는 과정에서
올바른 결과를 얻을 수 있습니다.

”



이 단원을 배우고 나면

여러분은 데이터를 효율적으로 관리하기 위해 데이터 압축과 암호화 및 빅데이터 기술을 활용하는 방법을 살펴보게 됩니다. 이를 바탕으로 데이터를 안전하게 관리하는 태도와 빅데이터를 분석하여 가치를 해석하는 능력을 기를 수 있습니다.



데이터

데이터의 압축과 암호화는 데이터를 효율적으로 관리하고 보호하는 데 도움을 줍니다. 또한 대량의 데이터를 효율적으로 관리하기 위해 윤리적 문제를 고려하여 빅데이터를 수집, 처리, 관리하는 과정을 경험해 봅니다.



협력적 창의·융합
프로젝트 소개

“안전 문제,
빅데이터로
해결해 볼까?”

84쪽에서
프로젝트 학습을
진행합니다.

“ 빅데이터를 활용한 K-Culture의 인기도 분석으로
한국의 다양한 문화를 전 세계 사람들에게 전달하고 있다.

”

K-Culture는 한류라고도 알려져 있으며,
한국의 음악, 드라마, 영화, 패션, 음식, 관광지 등 다양한 요소들을 포함합니다.
이러한 요소에 대한 검색량, 관련 게시물 등에서 데이터를 수집 및 분석하여
K-Culture 수요 동향을 파악할 수 있습니다.



우와! 멋지다!!
이걸 많은 사람들이
봐야 해!

REC

우리나라의
독특한 매력이
드러나는 콘텐츠가
인기네요.

우리나라의
전통 문화를
재미있게 알릴 수
있겠어요!

우리는 이처럼 다양한 데이터의 시대에 살고 있습니다.
데이터는 컴퓨팅 시스템을 활용한 문제 해결에 중요한 역할을 합니다.
따라서 데이터를 압축하고 암호화해서 안전하게 관리하고 보호할 줄
알아야 하며, 또 데이터를 수집, 분석, 시각화하여 문제를
해결하는 데 활용할 수 있어야 합니다.

이 단원의 핵심 역량

컴퓨팅 사고력

인공지능 소양

디지털 문화 소양

지식 정보 처리

창의적 사고

협력적 소통

공동체 역량

알고 있나요?



중학교 『정보』에서 배우요

디지털 데이터의 표현 방법

컴퓨터에서 처리할 수 있는 비트(0과 1)로 표현

문자

• 아스키 코드
• 유니코드

소리

• 표본화
• 양자화
• 부호화

이미지

• 화소(픽셀)
• 해상도

동영상

• 프레임
• 연속 재생

데이터의 수집과 분석

다양한 데이터를 목적에 맞게 수집·분석하여
융합적으로 문제 해결

데이터
수집

• 관찰
• 설문 조사
• 웹 검색
• 공공 데이터 활용 등

데이터
구조화

• 리스트
• 표(테이블)
• 트리
• 그래프 등

데이터
시각화

• 선그래프
• 원그래프
• 막대그래프
• 산점도 등

K-POP과 같은
한류 관련
검색량이 늘어!

BIG DATA

1

디지털 데이터의 압축과 암호화

우리는 스마트폰과 같은 전자 기기로 사진, 동영상, 통화 기록 등 다양한 데이터를 접한다. 그중 고화질 영상과 같은 데이터는 어떻게 볼 수 있을까? 내가 스마트폰에 저장한 사진은 어떤 기술이 적용되어 안전하게 보호될까? 바로 데이터를 압축하고 암호화하여 효율적이고 안전하게 관리하기 때문이다.



생각해 보세요?

우리가 데이터를 사용할 때 어느 정도의 에너지가 필요할까?



우리는 기후 변화의 원인이 되는 '온실 가스 배출'이라 하면 화력 발전 소나 자동차의 배기가스를 떠올린다. 그러나 전 세계 데이터 센터에서 배출하는 이산화탄소의 양은 2002년 기준 7,600만 톤에서 2020년 기준 2억 5,900만 톤으로 빠르게 증가하고 있다. 데이터를 저장하고 전송하는 과정에서 막대한 양의 에너지가 소모되기 때문이다. 우리가 별다른 생각 없이 쓰고 있는 데이터가 기후 변화에 큰 영향을 미친다는 것이다. '전기 먹는 하마'라 불리는 데이터 센터는 대규모 전력을 필요로 하고 이를 24시간 운영하기 위한 냉각 장치 가동으로 엄청난 전력 소모와 이산화탄소를 배출하고 있다.



디지털 데이터의 사용량 증가

사람들의 데이터 사용량은 데이터 무제한 요금제나 와이파이 서비스의 보편화로 지금도 폭발적으로 증가하고 있다.



이렇게 폭발적으로 증가하는 데이터를 효율적이고 안전하게 사용하기 위해서는 데이터의 저장 공간 축소, 전송 속도 향상, 데이터에 대한 접근 제어 등을 기억해야 한다.



단원
목표

데이터의 압축과 암호화의 필요성을 이해하고 데이터를 효율적으로 관리 및 보호할 수 있다.



핵심
질문

데이터를 효율적으로 관리하고 보호하는 방법은 무엇일까?

1-1 데이터의 압축

주요 용어

데이터 압축, 손실 압축, 무손실 압축

학습
목표

- 디지털 데이터 압축의 개념과 필요성을 설명할 수 있다.
- 간단한 데이터에 압축 기법을 적용하여 효율성을 분석할 수 있다.



일상생활에서 물건을 보관할 때 부피를 줄이기 위해 압축하는 경우가 있다. 이처럼 데이터를 전송하는 데 걸리는 시간을 줄이는 방법은 무엇일까?



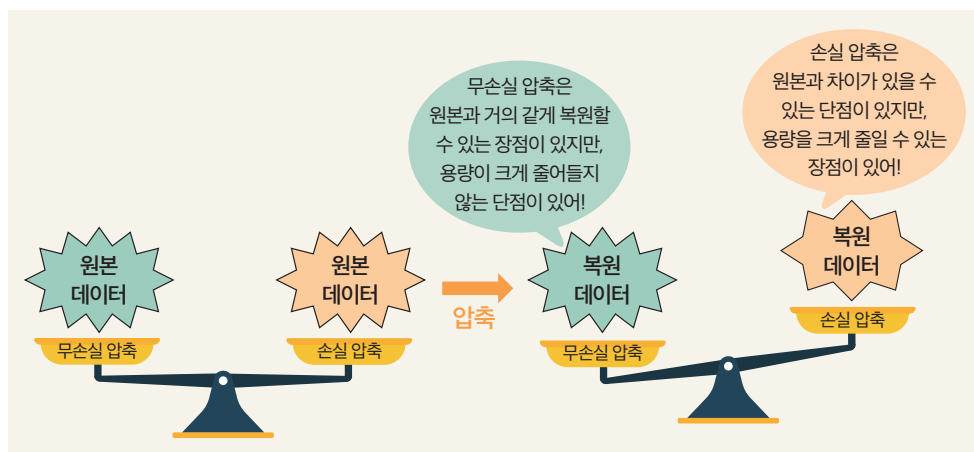
1 데이터 압축은 무엇일까?

우리는 일상생활에서 숫자, 문자, 이미지, 소리 등 다양한 형태로 정보를 표현한다. 컴퓨터는 여러 형태의 정보를 0과 1의 숫자로 변환하여 저장하는데 이처럼 일련의 숫자 형태로 변환한 정보를 디지털 데이터라고 한다. 디지털 데이터는 데이터를 저장하는 구조나 형태에 따라 정형 데이터와 비정형 데이터로 구분된다.

디지털 기술의 발달로 전 세계에서 많은 양의 디지털 데이터가 만들어지고 있다. 온라인으로 텍스트, 이미지, 영상 데이터를 공유하는 사람들이 많아지면서 디지털 데이터를 효율적으로 다루는 방법, 즉 데이터 압축에 대한 관심이 높아졌다. 데이터 압축은 원본 데이터를 더 작은 용량의 데이터로 만드는 것으로, 크게 무손실 압축 방법과 손실 압축 방법으로 나눌 수 있다.

정형 데이터와 비정형 데이터

- **정형 데이터**: 미리 정해 놓은 정해진 형태의 데이터로, 행과 열로 표현하는 테이블 등의 형태로 저장할 수 있다. 대표적인 예로 날짜, 이름, 주소 등이 있다.
- **비정형 데이터**: 형식이 정해지지 않은 데이터로, 문자, 소리, 이미지, 동영상 등이 대표적이다.



[그림 II-1] 데이터를 압축하는 방법 예시

무손실 압축은 데이터 손실 없이 압축을 하기 때문에 압축률이 낮아 용량이 큰 편이다. 텍스트로 된 파일, 은행 기록과 같은 데이터의 경우는 손실이 생겼을 때 문제가 될 수 있으므로 무손실 압축이 적합하다. 반면에 **손실 압축**은 원본 데이터의 손실이 일어나 압축 해제한 데이터와 원본 데이터 간 차이가 있을 수 있지만, 압축률이 높아 데이터 용량을 효과적으로 줄일 수 있다.

무손실 압축과 손실 압축

구분	무손실 압축	손실 압축
데이터 손실	손실 없음.	손실 있음.
원본과 비교	거의 같음.	다를 수 있음.
압축률	낮음.	높음.
용량	큰 편	작은 편
파일 형식	ZIP, GIF, PNG 등	JPEG, MPEG, MP3 등



디지털 기술의 발전으로 디지털 기기의 사용이 보편화되면서 디지털 데이터는 수정, 복제와 같은 데이터 가공 및 전송 작업이 편리해졌습니다.

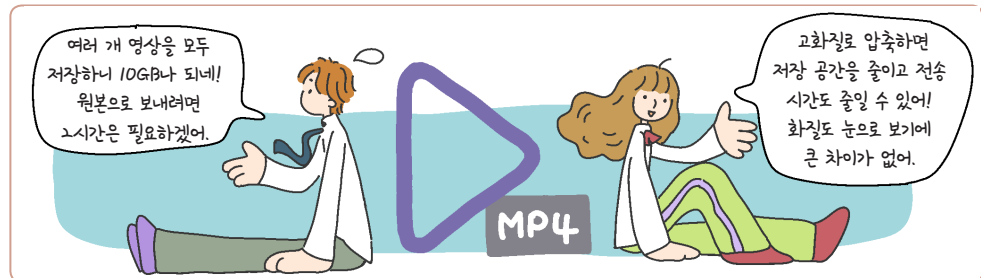
2 데이터 압축은 왜 필요할까?



다양한 분야에서 대규모 데이터를 활용하는 사례가 많아지면서 효율적인 데이터 관리와 전송의 중요성이 더욱 커지고 있습니다.

디지털 데이터는 현대 사회의 핵심 자원으로 개인뿐만 아니라 기업, 정부와 같은 다양한 조직에서 많은 양의 데이터를 생성한다. 이처럼 많은 양의 데이터를 압축하는 것은 데이터 저장 공간, 전송 시간을 줄이는 데 효과적이며 데이터 관련 비용을 낮추는 데 도움을 준다. 또한 대규모 데이터 관리, 전송에 필요한 전기 사용을 줄여 환경적인 측면에서도 이점이 있다. 데이터 보안의 측면에서도 압축된 데이터를 해독하려면 많은 비용이 필요하므로 데이터를 더 안전하게 관리할 수 있다.

오른쪽 내용은 데이터 압축 기술을 활용한 데이터 압축의 장점이다. 이를 통해 데이터 압축의 필요성을 알 수 있다.



데이터 압축 기술을 통해 고화질의 영상을 온라인 비디오 스트리밍 서비스로 감상하거나 실감나는 게임 그래픽을 즐길 수 있게 되었습니다.

데이터 압축 기술은 빠른 데이터 전송, 효과적인 데이터 관리를 가능하게 하여 다양한 분야에 영향을 미치고 있다. 예를 들어, 의료 분야에서는 의료 이미지, 환자 기록 데이터를 효과적으로 관리, 공유하기 위해 데이터를 압축한다. 이처럼 대규모 데이터셋을 압축하면 분석 결과를 더 빠르게 얻을 수 있고 데이터 관리 비용도 줄일 수 있다.

LINK 241쪽 부록 활동 2



함께 해결하기

데이터를 압축하는 이유

조사 실천

➔ 다음 내용을 읽고, 데이터 관리 측면에서 데이터 압축의 필요성을 알아보자.



탄소 중립은 배출한 탄소량만큼 탄소를 흡수해 실질 배출량을 0으로 만들기 위한 대책입니다.

기업은 많은 양의 데이터를 저장 및 관리하기 위해 데이터 센터가 필요하다. 그러나 데이터 센터를 유지하려면 엄청난 양의 전기가 필요하다는 문제점이 있다.

1. 데이터 센터에서 발생하는 이산화탄소를 줄이기 위해 실천하고 있는 친환경 데이터 센터 사례를 찾아보자.
2. 데이터 관리 측면에서 탄소 중립 실천을 위해 할 수 있는 일을 한 가지씩 적고 짝과 비교해 보자.

성취기준에 도달했는지 확인해 보세요.

성취도 평가

디지털 데이터 압축의 필요성을 설명할 수 있는가?

예 ☐ 아니요 ☐

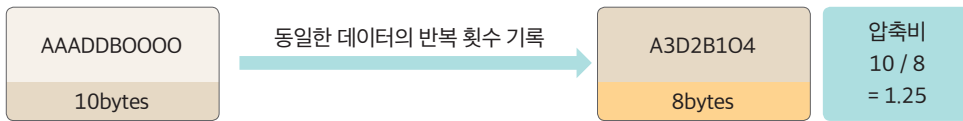
3 압축 효율성은 데이터 압축 방법에 따라 어떻게 다를까?

많은 양의 데이터가 급격히 늘어나면서 데이터 압축의 효율성이 중요해지고 있다.

1 무손실 압축 방법

데이터는 컴퓨터에 0과 1로 저장된다. 0과 1의 나열 순서를 기억해 두었다가 원본과 똑같이 적는 것이 바로 무손실 압축의 원리이다.

런 길이 부호화(Run length encoding) 동일한 데이터가 몇 번 반복되는지 기록하여 데이터를 짧게 만드는 방식이다. 런 길이 부호화를 사용한 대표적인 이미지 파일 형식으로 BMP가 있다.



Tip
0과 1의 나열을 어떻게 기억하느냐에 따라 압축 방법이 달라집니다.

LINK 152쪽 2번

압축비
압축 전 데이터의 크기를 압축 후 데이터의 크기로 나눈 값으로, 원본 데이터가 압축된 데이터보다 몇 배 더 큰지를 나타낸다.

런 길이 부호화는 아이콘이나 로고 등 단색을 사용한 이미지를 효율적으로 압축하지만, 다양한 색을 사용한 사진이나 디지털 이미지에는 부적합하다.

런 길이 부호화 압축은 연속적으로 반복되는 문자열을 감지하기 때문에 연속적인 반복이 많은 데이터에 효율적입니다.



소스로 해결하기

런 길이 부호화 방식

문자열을 이용하여 런 길이 부호화를 확인해 보자.

1. 문자열을 살펴보고 런 길이 부호화 방식에 따라 압축한 후, 원본 데이터 크기와 압축 데이터 크기, 압축비를 적어 보자.

문자열	압축 후 결과	원본 데이터 크기	압축 데이터 크기	압축비
NOOOOOOOOOO		11bytes	bytes	2.2
YEEEEEAH		bytes	8bytes	
HURRAY		bytes	10bytes	

2. 1에서 압축률이 낮게 나온 입력값을 토대로 런 길이 부호화 방식의 한계를 적어 보자.

3. 압축한 데이터를 복원해 보자.

압축 후 데이터 예	압축 전 데이터 예
H1A6P2Y1	HAAAAAPPY



성취기준에 도달했는지 확인해 보세요.

성취도 평가

데이터에 압축 기법을 적용하여 압축의 효율성을 분석하고 평가할 수 있는가?

예 ☐ 아니요 ☐

사전 압축 알고리즘

사전 자체가 공간을 차지하며, 압축한 데이터가 이 사전을 가지고 있어야 한다. 따라서 데이터가 한 번 이상 반복적으로 등장해야 공간을 절약할 수 있으며, 사전에 내용이 추가되는 만큼 많은 데이터를 압축할 수 있다. 따라서 중복되는 데이터가 많을수록 사전을 효율적으로 사용할 수 있다.

사전 압축(Dictionary compression)

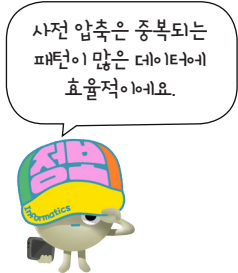
원본 데이터를 나열한 후 코드를 붙여 만든 사전을 바탕으로 압축하는 방식으로, 런 길이 부호화의 단점을 보완하였다.



허프만 부호화 알고리즘

1. 데이터에서 사용되는 각 문자에 대한 출현 빈도를 구한다.
2. 빈도를 기준으로 내림차순으로 정렬한다.
3. 값이 가장 작은 두 개의 문자를 묶고 0 또는 1 값을 할당한다. 값이 같다면 정렬 순서가 앞인 것에 0을 할당한다.
4. 아래에서 위로 트리를 형성한다.
5. 위에서 아래로 코드를 조합한다.

가장 대표적인 사전 압축 방식은 기본 사전 압축 방식을 개선한 **허프만 부호화(Huffman coding)**이다. 허프만 부호화는 전체 데이터에서 자주 등장하는 값별로 사용 빈도를 계산하여 사용 빈도가 높을수록 비트 값을 적게 차지하는 짧은 이진수를 부여하는 방법이다. 상대적으로 적은 용량으로 데이터를 저장할 수 있으므로 기본 사전식 압축 방식보다 효율이 높다.



[그림 Ⅱ-2] 허프만 부호화 방식



지식 키우기

허프만 코드

‘AAADDBOOOO’ 문자열을 허프만 부호화해 볼까?

① 문자 빈도 분석하기

‘A’는 3번, ‘D’는 2번, ‘B’는 1번, ‘O’는 4번 등장한다. 이를 내림차순으로 배열한다.

② 빈도 기반 이진 트리 만들기

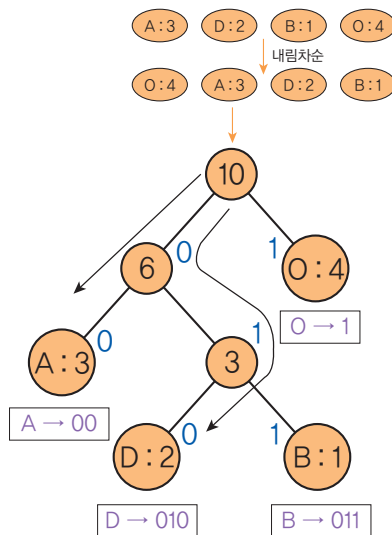
허프만 트리는 문자 빈도를 기반으로 아래에서 위로 트리를 형성한다. 이때 생성하는 가지는 최대 2개씩 이진 트리로 구성하는데 주어진 문자열에서 등장하는 순서대로 왼쪽 또는 오른쪽으로 나열하여 오른쪽 그림처럼 트리를 구성할 수 있다.

③ 이진 부호 붙이기

왼쪽으로 이동할 때는 ‘0’, 오른쪽으로 이동할 때는 ‘1’을 추가하여 ‘O’:1, ‘A’:00, ‘B’:011, ‘D’:010으로 이진 부호를 붙일 수 있다.

④ 데이터 변환하기

문자열 ‘AAADDBOOOO’를 허프만 부호로 변환하면 ‘0000000100 10011111’이 된다.



2 손실 압축 방법

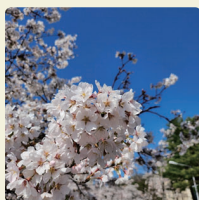
정보 과학 기술의 발달로 우리는 고품질의 이미지 데이터 또는 영상 데이터를 쉽게 접할 수 있게 되었다. 이미지 데이터는 품질이 높아질수록 용량이 커지며, 소리 데이터는 음질이 높을수록 용량이 커진다. 이를 효과적으로 저장하고 전달하며 관리하기 위해 사람의 눈으로 보거나 귀로 들었을 때 원본 데이터와 복원 데이터의 차이가 크지 않으면서 용량을 획기적으로 줄이는 손실 압축 기술이 발달하게 되었다.

이미지 압축 방식

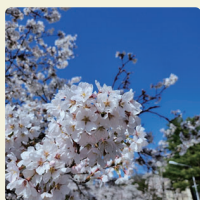
JPEG는 대표적인 손실 압축 방법이다.

JPEG 압축 방식은 사람의 눈이 색의 차이보다 밝기 차이를 더 잘 인식한다는 점과 사진을 이루는 개별 픽셀의 색은 주변에 있는 픽셀의 색과 비슷하다는 성질에 기반을 둔다.

원본



색 변화



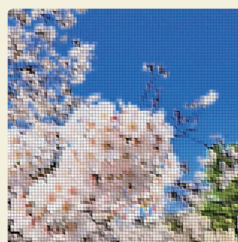
밝기 변화



사람의 눈은 색보다 밝기 차이를 더 잘 인식한다.



각 픽셀의 R(빨강), G(초록), B(파랑) 값을 모두 저장하지 않고, 픽셀의 휘도, 파란색 차이, 빨간색 차이 정보를 바탕으로 압축한다.



개별 픽셀의 색은 주변에 있는 픽셀의 색과 비슷하다.



이미지를 8×8 픽셀 블록으로 나누어 압축한다.

JPEG(Joint Photographic Experts Group)

- 압축률을 높이기 위해 사람이 잘 인식할 수 없는 부분의 정보를 변형하거나 제거하여 용량을 줄이는 손실 압축 방법 중 하나이다.
- 컴퓨터나 디지털카메라에서 사진을 표현할 때 많이 사용하며, JPG 확장자로 나타낸다.

휘도(Luminance)

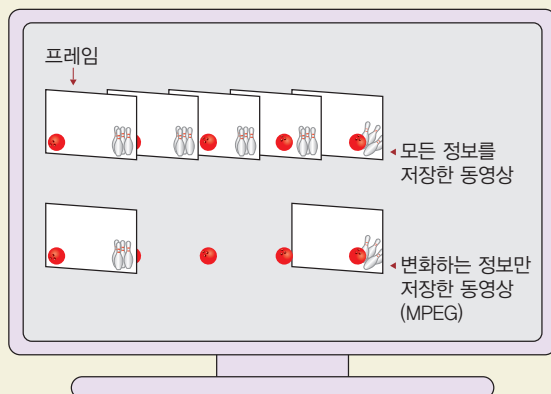
빛의 정도(밝기)를 의미한다.

영상 압축 방식

MPEG은 대표적인 영상 압축 방식이다.

영상 데이터는 여러 장의 이미지와 소리 데이터로 이루어져 용량이 무척 크다. MPEG은 기본적으로 JPEG 압축 방법을 사용하고, 연속된 이미지의 유사성을 이용한다. 각 프레임이 서로 다르지만 연속된 화면에서 변화하는 부분을 찾아내어 차이만 저장한다.

영상을 구성하는 각 이미지를 프레임이라고 한다. 앞서 살펴본 JPEG로 영상의 프레임을 압축할 수도 있지만, 고품질 영상을 압축하려면 더 효율적인 압축 기술이 필요한 것이다.



MPEG(Moving/Motion Picture Experts Group)

- JPEG와는 달리 시간에 따라 연속적으로 변하는 동영상이나 오디오 데이터를 압축하고 해제하는 손실 압축 방법 중 하나이다.
- 동영상이나 오디오는 프레임 데이터를 사용하여 변환하고 압축하는 방식으로 mp3, mp4 등 확장자로 나타낸다.



표본화, 양자화, 부호화

- **표본화**: 아날로그 신호를 디지털로 변환하기 위해 일정한 시간 간격으로 진동을 측정한다.
- **양자화**: 표본값을 일정한 단계로 나누어 디지털값으로 변환한다.
- **부호화**: 변환된 디지털값을 비트(0과 1)로 표현한다.

궁금해요!

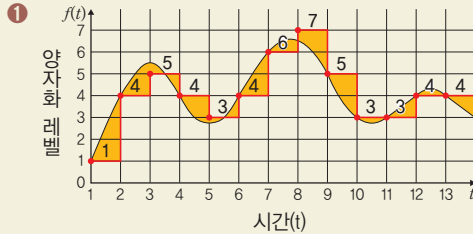
소리 정보를 손실시키지 않고 용량만 줄일 순 없을까?

소리 데이터도 무손실 압축이 가능하다. 손실 압축에 비해 압축률은 낮지만 비교적 높은 품질의 음원을 저장할 수 있다. 대표적인 포맷으로 FLAC, ALAC가 있다.

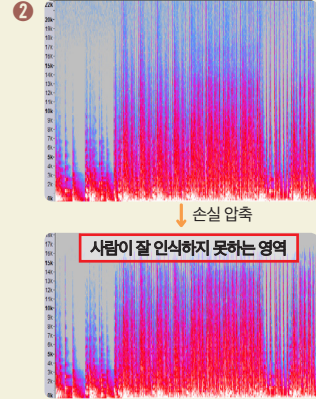
소리 압축 방식

아날로그 소리는 **표본화, 양자화, 부호화**를 거쳐 디지털 신호로 기록된다.

- 1 양자화 과정에서 소리의 표본값을 일정한 단계로 나누어 디지털값으로 변환하므로 실제값과 양자화된 값 사이에 차이가 생긴다.
- 2 더 효율적으로 압축하기 위해 사람이 잘 인식하지 못하는 부분의 정보는 손실시키고 압축한다.



부호화
1 4 5 4 3 4 6 7 5 3 3 4 4
↓
001 100 101 100 011 100 110 111 101 011 011 100 100
곡선은 아날로그 값(실제값), 빨간색 직선은 양자화된 값을 의미한다. 노란색 영역만큼 차이가 존재한다.



이러한 원리를 바탕으로 압축된 데이터는 일상생활에서 다양하게 사용된다. 특히 빅데이터가 많이 활용되면서 데이터를 효율적으로 관리하기 위한 압축 기술의 중요성이 높아지고 있다. 미래에는 현실에 가까운 형태의 고품질 데이터가 많아짐에 따라 압축 기술이 중요해지는 만큼 탄소 배출량을 줄이는 친환경적인 데이터 관리에도 많은 관심을 둘 필요가 있다.



스스로 해결하기

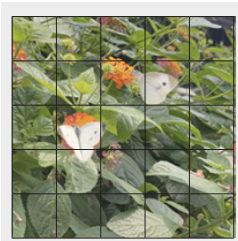
영상 압축의 원리

분석 실습

➡ 다음은 영상의 연속된 프레임을 나열한 것이다. 컴퓨터를 이용해서 영상을 압축하고 원리를 확인해 보자.

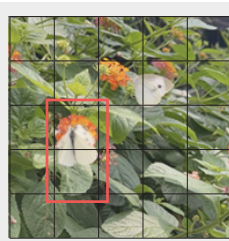
- 1 첫 번째 프레임을 기준으로 두 번째 프레임에서 달라진 부분을 참고하여, 세 번째 프레임에서 달라진 부분의 격자 개수를 세어 보자.

첫 번째 프레임



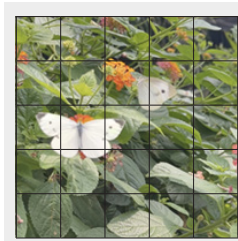
달라진 부분의 격자 개수

두 번째 프레임



2개

세 번째 프레임



눈으로 변화를 확인할 수 있는 부분의 격자만 세고, 각 프레임의 격자 1개의 용량은 10bytes라고 가정합니다.

- 2 3개의 프레임 이미지를 모두 저장할 때와 달라진 격자 부분만 저장할 때 필요한 저장 공간의 크기를 계산해 보자.

모든 격자를 저장할 때 필요한 저장 공간	달라진 부분의 격자만 저장할 때 필요한 저장 공간
10bytes × 25개 × 3장 = 750bytes	10bytes × (25개 + 2개 + □개) = □bytes

성취기준에 도달했는지 확인해 보세요.

성취도 평가

기본적인 압축 원리를 이해하고 간단한 데이터에 적용할 수 있는가?

예 ☐ 아니요 ☐



➔ 다음 실습을 통해 이미지 무손실 압축 방법을 비교해 보자.

실습 자료



PIL(Python Imaging Library)은 이미지 분석 및 처리를 쉽게 할 수 있는 라이브러리로, 다양한 이미지 파일 형식을 지원합니다.



PNG(Portable Network Graphics)는 그림의 정보를 변형하지 않고 용량만 줄이는 무손실 압축 방법 중 하나로, 인터넷에서 고 품질의 정밀한 그림을 표현할 때 주로 사용됩니다.



GIF(Graphics Interchange Format)는 무손실 압축 방법 중 하나로, 256색만 표현 가능하여 그림의 품질은 낮지만, 여러 장의 그림을 넣고 차례대로 보여 주는 애니메이션 기능이 있습니다.

- 1 사용할 BMP 이미지(.bmp)를 준비한 후 코랩을 열고 이미지 변환에 필요한 라이브러리를 불러온다.

```
import os
from PIL import Image
```

- 2 BMP 이미지의 해상도, 파일 크기를 확인한다. (BMP 이미지 파일 업로드 방법은 QR코드에서 확인한다.)

```
im=Image.open('flower_BMP.bmp')
print('이미지 확장자', im.format)
print('이미지 해상도', im.size)
print('파일 크기', os.path.getsize('flower.bmp.bmp'), 'Bytes')
```

실행 결과

- 3 같은 이미지를 PNG, GIF 이미지로 변환하여 해상도와 파일 크기를 확인한다.

```
im.save('flower_PNG.png', 'png')
im2=Image.open('flower_PNG.png')
print('이미지 확장자', im2.format)
print('이미지 해상도', im2.size)
print('파일 크기', os.path.getsize('flower_PNG.png'), 'Bytes')
im.save('flower_GIF.gif', 'gif')
im3=Image.open('flower_GIF.gif')
print('이미지 확장자', im3.format)
print('이미지 해상도', im3.size)
print('파일 크기', os.path.getsize('flower_GIF.gif'), 'Bytes')
```

실행 결과

- 4 BMP, PNG, GIF 이미지를 출력하여 눈으로 차이를 구분할 수 있는지 확인해 본다.

	bmp	png	gif
im im2 im3			

- 5 세 가지 압축 방법을 비교해 보고, 어떤 압축 방법이 가장 효율적일지 서로 이야기 나누어 보자.



성취기준에 도달했는지 확인해 보세요.

성취도 평가

원본과 압축 데이터의 품질 및 용량 분석, 압축 방법 간의 효율성을 평가할 수 있는가?

예 ☐ 아니요 ☐



개념 정리하기

오른쪽 주요 용어에서 알맞은 것을 골라 써넣으시오.

- ☒ _____ 은 원본 데이터를 더 작은 용량의 데이터로 만드는 것을 말한다.
- ☒ 데이터는 압축하는 과정에서 데이터의 손실 유무에 따라 데이터가 손실되지 않는 _____ 과 데이터가 손실되는 _____ 으로 나뉜다.

주요 용어

데이터 압축

손실 압축

무손실 압축

1-2 데이터의 암호화

주요 용어

암호화, 복호화, 치환형 암호화, 전치형 암호화, 단방향 암호화, 대칭 암호화, 비대칭 암호화

학습
목표

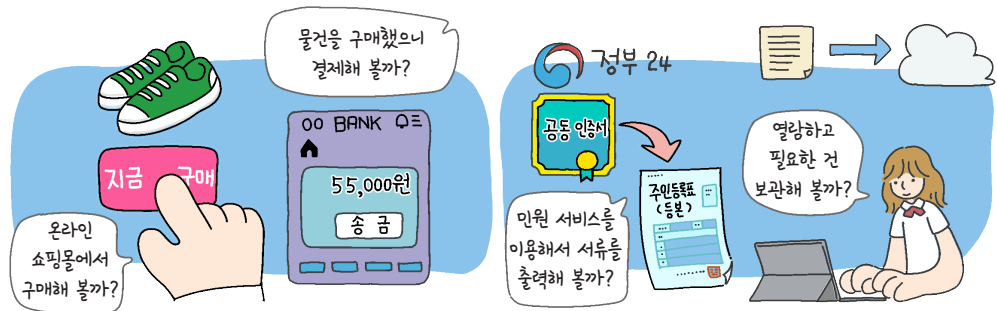
- 암호화의 개념과 중요성을 설명할 수 있다.
- 암호화를 활용하여 데이터를 효율적이고 안전하게 관리하는 사례를 비교 및 분석할 수 있다.

? 누구에게도 들리지 않고 문서를 전달해야 하는데 문서를 들키고 말았다.
문서가 발각되어도 내용을 들키지 않게 숨길 수 있는 방법이 있을까?



1 암호화란 무엇이며, 암호화는 왜 중요할까?

인터넷이 발전하면서 온라인 쇼핑, 민원 서류 발급 등 다양한 서비스가 네트워크 상에서 이뤄지게 되었다.



암호(cryptography)라는 단어는 그리스어로 비밀이라는 뜻의 크립토스(Kryptos)에서 기원하였습니다.

우리가 이와 같은 서비스를 이용할 수 있는 원리는 무엇일까? 바로 중요한 정보(비밀번호)를 다른 사람은 알지 못하게 전달하는 기술 덕분이다. 이처럼 정보를 읽기 어려운 형태로 바꾸거나 반대로 다시 읽을 수 있게 해독하는 기술을 암호화(encryption)라고 한다.

암호화의 기능

- **기밀성(confidentiality)**: 정보를 보호하는 속성으로 권한이 있는 사용자만 정보를 볼 수 있다.
- **무결성(integrity)**: 정보의 내용이 임의로 생성, 변경, 삭제가 일어나지 않았는지 확인한다. 즉, 원본과 같은지 확인할 수 있다.
- **인증(authentication)**: 권한이 있는 사용자인지 식별하는 것이다. 예를 들어, 계좌에 대한 접근이 실제로 계좌의 주인에게서 온 것인지 확인한다.



[그림 11-3] 암호화의 기본 원리

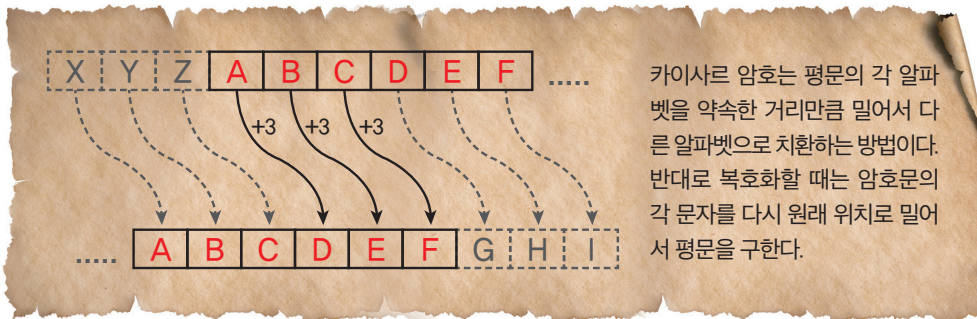
암호화를 통해 보호하고자 하는 원본 데이터를 **평문**(plaintext)이라고 하며, 암호화된 데이터는 **암호문**(ciphertext)이라고 한다. 평문을 암호문으로 변환하는 과정을 **암호화**라고 하며, 암호문을 평문으로 해독하는 과정을 **복호화**(decryption)라고 한다. 암호화와 복호화에 필요한 열쇠가 되는 규칙은 **암호키**(encryption key)라고 한다.

2 암호화 방법에는 어떤 것이 있을까?

고대 사회에서는 비밀을 유지하며 정보를 전달하기 위해 다음과 같은 방법을 사용하였다. 살아 있는 동물의 털을 깎아 메시지를 작성한 후 털을 길러 보내면 받는 사람이 털을 깎아 확인하는 방법, 받았을 때는 보이지 않지만 불빛에 반응하는 약품으로 메시지를 작성해서 보내는 방법 등을 들 수 있다. 또 어떤 방법이 있었을까?

1 치환형 암호화 방법

치환형 암호화는 평문의 각 문자를 약속된 다른 문자로 **치환**하는 암호화 방법이다. 복호화를 위해 암호문을 주고받는 발신자와 수신자는 각 문자를 어떤 문자로 치환할지 약속해야 한다. 암호학에서 다루는 간단한 예로, 로마 황제 ‘율리우스 카이사르’의 이름을 딴 카이사르 암호(caesar cipher)가 있는데, 그 방법은 다음과 같다.



[그림 II-4] 키가 3일 때 카이사르 암호화

치환(substitution)

특정 부분을 다른 문자로 바꾸는 것을 말한다.

궁금해요!

카이사르 암호는 어떻게 만들어졌을까?

이 암호는 ‘율리우스 카이사르’가 전쟁 중 군사 정보를 비밀스럽게 전달하기 위해 고안했다고 전해지는데, 문에 카이사르 암호라고 불리게 되었다. 이 암호는 알파벳에 숫자를 배정하고, 숫자 순서대로 알파벳을 정렬하여 암호화하는 방식으로, 암호화 및 복호화가 간단하다.



소스로 해결하기

치환형 암호화 방법

분석

실습

➔ 치환형 암호화를 만들어 보고, 치환형 암호화의 한계점을 알아보자.

1. 카이사르 암호 방법으로 암호화의 키가 3일 때 아래의 암호문을 해독해 보자.

PHHW DW VFKRRO DW ILYH SP
암호문



M
평문

2. 아래 암호문을 해독해 보고 한계점에 대해 생각해 보자.

FIEYXC FIKMRW XLI QSQIRX CSY HIGMHI XS FI CSYVWIPJ. FI CSYVWIPJ, RS SRI GER WEC CSYVI HSMRK MX AVSRK.
암호문



BEAUTY BEGINS THE MOMENT
평문



1번은 암호문의 각 영문자에서 왼쪽으로 3(키)만큼 이동한 영문자를 찾고, 2번은 주어진 평문에서 키를 찾아 정답을 채워 봅니다.



성취기준에 도달했는지 확인해 보세요.

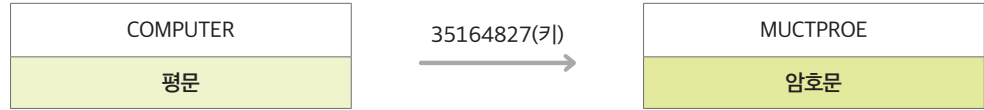
성취도 평가

치환형 암호화 방법과 그 효율성을 설명할 수 있는가?

예 ☐ 아니요 ☐

2 전치형 암호화 방법

전치형 암호화 방법은 문자의 순서를 바꾸어 암호화하는 방법이다. 암호문을 주고 받는 발신자와 수신자는 평문의 문자를 재배치할 순서를 키로 나누어 갖는다. 단순 전치 암호가 대표적인 예로서, 평문을 키의 재배열 순서에 따라 재배치한다.



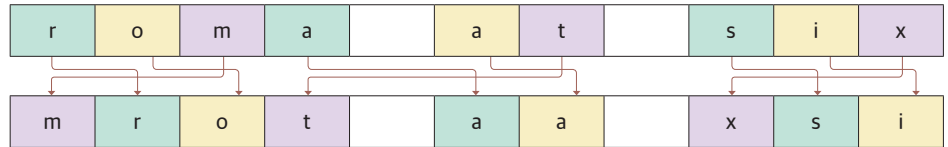
평문이 길다면 키의 길이만큼 일정 간격으로 나눈 후, 나눈 문자를 키의 재배열 순서에 따라 재배치하는 방법도 있다.

〈예제 1〉의 암호화한 결과를 다시 복호화하여 평문을 만들어 보세요.



예제 1 | 'roma at six'라는 평문을 312라는 키로 암호화해 보자.

평문을 세 문자로 나눈 후 키의 순서에 따라 재배열하면 'mrot aa xsi'가 된다. 반대로 복호화할 때는 231 순서로 재배열하면 평문으로 복호화할 수 있다. 이때 키 순서에서 빈칸은 적용하지 않는다.



스키테일(scytale)

양피지를 나무 막대에 말아서 사용하였다. 원통에 감아진 텍스트가 일렬로 보이지만, 양피지를 펼치거나 지름이 다른 원통에 감으면 해독할 수 없다.



예제 2 | 고대 전쟁에서 그리스 군사들이 사용한 '스키테일' 암호문을 살펴보자.

특정 지름의 막대기에 종이를 감고 메시지를 작성한 후 같은 종이를 풀면 각 문자가 재배치된다. 이는 암호문 송신자와 같은 지름의 막대기를 가진 사람만 암호문을 해석할 수 있다. 이런 방법으로 그리스인들은 전쟁 중 군령 및 각 임무를 안전하게 전달할 수 있었다.

그러나 전치형 암호화는 암호를 해독할 수 있으므로 보안에 취약하다. 특히 컴퓨팅 기술이 발전하면서 별도의 안전한 암호화 방법이 필요하게 되었다.

스스로 해결하기

전치형 암호화 방법

분석 실습

➡ 전치형 암호화 방법에서 익힌 두 개의 암호화 방법을 이용하여 메시지를 해독해 보자.

1. 단순 전치 암호에 키 213으로 평문을 암호화한 암호문을 해독해 보자.



2. 아래 암호문은 스키테일 원통에 작성된 것이다. 한계점에 대해 생각해 보자.

암독세호해요를보!해

공백은 무시하고 스키테일의 원리를 찾아 해결해 봅니다.

성취기준에 도달했는지 확인해 보세요.

성취도 평가

전치형 암호화 방법과 그 효율성을 설명할 수 있는가?

예 ☐ 아니요 ☐

3 현재에는 암호화가 어떻게 사용되고 있을까?

정보 통신 기술이 발달함에 따라 인터넷을 통해 전달하는 데이터의 양은 급격하게 증가하였고, 이를 보호하기 위한 암호화 기술의 사용도 증가하고 있다.

1 단방향 암호화

암호화는 가능하지만 복호화는 불가능한 방식이다. 암호화만 가능하기 때문에 단방향 암호화라고 한다. 단방향 암호화는 동일한 평문에 대해 동일한 암호문을 생성하고, 한 글자라도 다른 평문은 다른 암호문을 생성하는 특징이 있기 때문에 다음과 같이 암호화된 값을 서로 비교해 무결성이나 인증에 사용된다.



단방향 암호화를 적용한 사례로는 사용자의 신원을 확인하고 전염병의 경로를 파악하는 QR코드가 있다.



QR코드로 신원 인증에 사용되는 단방향 암호화

COVID-19 때 전염 경로를 파악하고, 확산을 방지하기 위해 QR코드로 신원을 인증하는 전자 출입 명부를 시행하였다. 사용자가 서버에 QR코드 생성을 요청하면 서버에서 사용자의 정보를 포함한 암호문으로 QR코드를 생성한다. 인증 제한 시간 안에 QR코드를 찍으면 인증이 완료된다.

개인용 컴퓨터와 인터넷 상거래가 발달하면서 이전에는 군 사용이나 외교용으로 사용되었던 암호가 개인 또는 기업에서도 활용하게 되었습니다.

암호화의 기법 종류

- 단방향 암호화
- 양방향 암호화
 - 대칭 암호화
 - 비대칭 암호화

궁금해요!

해시 함수와 해시값은 무엇일까?

해시 함수는 다양한 길이의 입력 정보를 고정된 길이의 해시값으로 변환하는 함수이다. 이 과정은 단방향으로 한 번 변환된 데이터는 원래 상태로 복구할 수 없기 때문에 단방향 암호화라고 한다. 이 특성 때문에 해시 함수는 비밀번호와 같은 중요 정보를 안전하게 저장하는 데 사용된다. 사용자가 비밀번호를 입력하면 컴퓨터는 이를 해시값으로 변환해 저장하고, 나중에 같은 비밀번호를 입력했을 때 저장된 해시값과 비교해 일치 여부를 확인한다. 해시값은 입력값이 조금만 바뀌어도 완전히 다른 결과값이 나오기 때문에 해시값으로는 원래 입력값을 예측할 수 없다. SHA-256은 해시 함수의 한 예로, 비트코인과 같은 블록체인에서 거래 정보를 보호하는 데 사용된다.



스스로 해결하기

단방향 암호화 방법

분석 실습

➡ 단방향 암호화인 해시 함수를 사용해서 메시지를 암호화하고 확인해 보자.

1. 해시 함수 실습 웹 사이트(<http://isweb.joongbu.ac.kr/~miss/H.P/hash.html>)에 메시지를 입력하고,

SHA256 알고리즘으로 단방향 암호화를 한 결과인 해시값을 확인해 보자.

이 글은 원본입니다. ⇨ db910854d7a026c5dda4a27285239d63c1708bae4df634d56d157b8475658c69

2. 메시지를 아주 조금만 수정하여 다시 단방향 암호화를 한 결과를 원본과 비교해 보자.

이 글은 원본입니다 ⇨ 92369f694193bb2035d34f28420454837f20419252710d5b4170cd3cd71914f0

마침표 하나만 수정해도 전혀 다른 결과가 나온다는 사실을 알 수 있습니다.



성취기준에 도달했는지 확인해 보세요.

성취도 평가

단방향 암호화 방법과 그 효율성을 설명할 수 있는가?

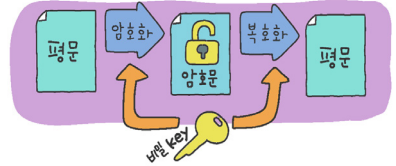
예 ☐ 아니요 ☐

2 대칭 암호화

대칭 암호화는 양방향 암호화 종류로 송수신자가 같은 키를 이용해서 암호화와 복호화를 하는 방법이다. 송수신자가 같은 키를 공유하거나 정보를 공유할 사람들만 키를 공유하므로 비밀키라고 부른다.

대칭 암호화는 다음과 같은 특징이 있다.

- 대칭 암호화는 암호화 속도가 빠르다.
- 암호문을 도둑맞아도 비밀키가 없으면 복호화할 수 없다.
- 대칭 암호화는 키를 교환하는 문제(키 배송 문제)가 발생한다. 특정한 사람에게 암호화된 정보와 키를 함께 전달하는 과정에서 분실이 생길 수 있기 때문이다.
- 개인마다 키를 개별적으로 교환해야 하므로 관리해야 할 키가 방대하다. 사용자가 10만 명이라면 10만 개의 키를 보관해야 한다.



대칭 암호화를 사용할 때 발생하는 문제는 무엇일까요?



대칭 암호화를 적용한 사례로는 인터넷 뱅킹에 사용되는 OTP와 보안 카드가 있다.

OTP(One Time Password)

의뢰인	홍길동
출금 계좌	12345678901234
승인 전화 번호	02-12345678
일회용 비밀번호	123456

OTP를 제작할 때 기기와 서버에 비밀 키를 할당하고, OTP 기기의 버튼을 누르면 현재 시간과 비밀키를 이용해 생성한 번호로 사용자를 인증한다.

일회용 비밀번호인 OTP와 보안 카드는 공유된 비밀 키를 이용한다. 사용자의 신원을 인증하는 대표적인 대칭 암호화 사례이다.

보안 카드

0000은행		No. 0123456789		
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25
26	27	28	29	30
31	32	33	34	35

이 카드는 타인에게 양도·대여 할 수 없으며, 본인의 관리소홀로 발생한 손해는 은행이 책임지지 않습니다. 또한 이 카드를 분실 또는 타인에게 노출시킨 경우 즉시 신고해 주시기 바랍니다. 전화: 1234-5678

은행마다 1인당 한 장의 보안 카드를 발급한다. 번호가 정해져 있으며, 그때 그때 지정된 번호를 입력해 사용자를 인증한다.



함께 해결하기

대칭 암호화 방법

실습 토의

➡ 한컴오피스 프로그램을 이용해서 대칭 암호화를 확인해 보자.

1. 문서를 작성하고 권한이 있는 사용자만 문서를 확인할 수 있도록 다음 과정에 따라 암호를 설정해 보자.

- ① 한컴오피스로 한글 파일을 생성한다.
- ② 도구창에서 [보안] - [문서 암호 설정] - [열기/쓰기 암호]에 각각 암호를 입력한다.
- ③ 문서 저장 후, 다시 문서를 열려면 오른쪽 창에 암호를 입력해야만 한다.

열기/쓰기 암호 설정

열기 암호	설정(D)
새 암호	취소
암호 확인	변경/해제(E)
쓰기 암호	
새 암호	
암호 확인	

암호는 5자 이상 입력하세요.
암호를 잊어버리면 문서를 불러올 수 없습니다.

2. 대칭 암호화를 사용하는 이유와 그 중요성에 대해 서로 이야기 나누어 보자.

성취도 평가 대칭 암호화를 사용하는 이유와 그 안전성을 설명할 수 있는가?

예 ☐ 아니요 ☐

성취기준에 도달했는지 확인해 보세요.

3 비대칭 암호화

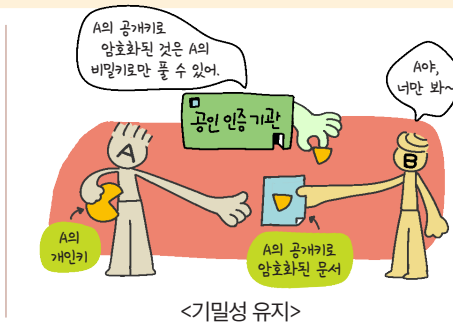
개인마다 키를 개별적으로 교환해야 하는 대칭 암호화의 문제를 해결하기 위해 키를 교환하지 않는 양방향 암호화의 또다른 종류인 비대칭 암호화 방식이 탄생하였다. 비대칭 암호화는 송수신자가 서로 다른 키를 갖는다. 송신자는 개인키(또는 공개키)로 암호화를 하고, 수신자는 개인키와 쌍을 이루는 공개키(또는 개인키)로 복호화한다.

비대칭 암호화를 적용한 사례로는 사용자의 신원을 확인하는 전자 서명과 비대칭 암호화 원리를 사용하고 있는 공동 인증서가 있다.

전자 서명은 이용자의 '개인키'를 사용하여 생성한 전자 서명을 전송함으로써 수신자에게 현재의 이용자가 본인임을 확인하는 과정에 사용된다. 이때 개인키와 공개키를 생성, 발급, 관리하는 공인 인증기관(CA: Certificate Authority)이 필요하다.



<전자 서명>



<기밀성 유지>

개인키는 도장과 같은 개념으로 항상 사용자가 소유해야 하고 타인에게 노출되지 않도록 모바일, PC 등 개인적인 공간에서 관리해야 합니다.

비대칭 암호화의 특징

비대칭 암호화 방식은 처리 속도가 느리지만 키 교환이 쉽다는 장점이 있다.

문서를 암호화하기 위해 공개키를 사용하는 것은 문서에 대한 접근을 통제하고 기밀성을 유지하는 데 도움이 돼요.



LINK 230쪽 전자 서명



함께 해결하기

비대칭 암호화 방법

분석

➔ 대칭 암호화와 비대칭 암호화를 비교하여 표를 완성해 보자.

구분	대칭 암호화	비대칭 암호화
키 관계	암호화 키와 복호화 키가 같음.	암호화 키와 복호화 키가 다름.
키 전송 여부		
장점	구현이 쉬움. 인증 속도가 빠름.	
단점		해독 시간이 오래 걸림.
활용	OTP, 보안 카드	공동 인증서



성취기준에 도달했는지 확인해 보세요.

성취도 평가

비대칭 암호화의 중요성과 필요성을 설명할 수 있는가?

예 ☐ 아니요 ☐



개념 정리하기

오른쪽 주요 용어에서 알맞은 것을 골라 써넣으시오.

- ☒ 암호문을 변환하는 과정을 , 해독하는 과정을 라고 한다.
- ☒ 평문의 각 문자를 약속된 다른 문자로 바꾸는 와 문자의 순서를 바꾸는 방법이 있다.
- ☒ 현재 사용하는 암호화 방법은 , , 가 있다.

주요 용어

암호화	복호화
치환형 암호화	전치형 암호화
단방향 암호화	대칭 암호화
비대칭 암호화	



디지털 데이터의 압축과 암호화 정리



데이터 관리는 똑똑하게!

➔ 다음 활동을 수행하면서 디지털 데이터의 압축과 암호화의 개념을 정리하고, 데이터를 효율적으로 관리하고 보호하는 방안을 알아보자.

📌 지식·이해

1. 디지털 데이터의 압축과 암호화의 개념을 중요성 또는 필요성 측면에서 적어 보자.

📌 과정·기능

2. 다음 문자열을 조건에 맞춰 암호화하고 압축해 보자.

<조건>

- 단순한 전치형 암호화이며 키는 4213이다.
- 사전식 압축 방법으로 압축하며 등장하는 순서대로 사전에 구성한다.

<원본 데이터>

BIBIDIBABIDIBU

<암호화>

14bytes

bits

압축

데이터	코드
	001
B	010
	011
D	100
	101

001 001 010 010
011 001 100
010 001 001 010
100 101 010

bits

압축비

/

≈ 2.67



54쪽, 60쪽을 확인하고 조건에 맞춰 빈 칸을 완성해 보세요.

📌 가치·태도

3. 환경을 고려하는 지속 가능한 발전의 측면에서 구체적인 데이터 사용 실천 방법 두 가지를 적어 보자.

• 실천 방법 1:

• 실천 방법 2:

📌 학업 성취도 점검!

평가 기준을 읽고, 스스로 자신의 실력을 점검해 보세요.



3개 활동을 마쳤으면 학업 성취도를 점검한 후, 보충 또는 심화 활동을 해 보세요.

내용 요소	평가 기준	성취도	
		예	아니요
지식·이해	디지털 데이터 압축과 암호화의 중요성과 필요성을 설명할 수 있는가?	65쪽 '심화' 활동으로 가시오.	65쪽 '보충' 활동으로 가시오.
과정·기능	디지털 데이터의 압축과 암호화 방법을 이해하고 설명할 수 있는가?		
가치·태도	데이터를 안전하게 관리하고 보호하는 태도를 지닐 수 있는가?		

부족한 부분은 채우면서
배움의 깊이를 더해 봅시다.



선택 활동

☒ 다음 선택 활동 중 나에게 알맞은 활동을 골라 해결해 보자.



암호는 다른 사람이 보더라도 그 의미를 알 수 없도록 만들어야 한다. 나만의 한글 암호화 규칙을 만들고 원하는 문구를 암호화해 보자.

예 버디의 암호화 규칙: 각 글자 뒤에 자음 비읍을 넣고 앞 글자 모음을 따라 쓰되, 받침은 끌고 온다.

- 원본 문장: 안녕하세요
- 암호 문장: 아반녀병하바세베요보

_____의 암호화 규칙

• 원본 문장 _____

• 암호 문장 _____



인공지능을 활용한 압축 기술에 대한 관심이 높아지고 있다. 코덱(CODEC)은 음성 또는 영상 데이터를 디지털로 변환해 저장하고 재생하는 소프트웨어 또는 하드웨어이다. 인공지능 기반의 압축 기술이 기존 코덱보다 좋은 점을 조사하여 정리해 보자.



압축된 데이터를 친환경적으로 관리하는 방법을 찾아보세요.



압축 기술

기존 코덱보다 좋은 점

예 인공지능을 이용한 영상 압축 기술

학업 계획 점검!



‘디지털 데이터의 압축과 암호화’ 단원을 학습한 후, 50쪽 핵심 질문의 답과 새롭게 알게 된 점을 적어 보세요.

핵심 질문에 답하기

새롭게 알게 된 점

2

빅데이터

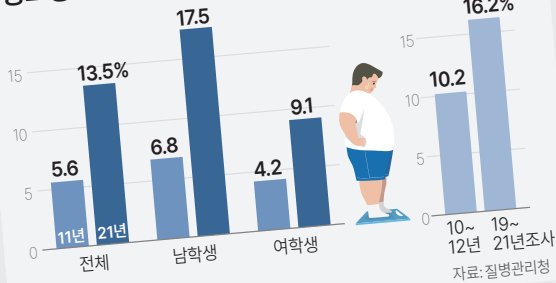
빅데이터는 합리적 의사 결정을 위한 유용한 도구로 데이터 분석을 통해 문제를 해결하는 데 활용한다. 빅데이터를 활용한 데이터 분석은 데이터를 수집하고 처리 및 관리하는 과정이다. 이 과정에서 윤리적인 문제를 고려해서 수행해야 올바른 결과를 도출할 수 있다.



생각해 보세요?

우리가 필요로 하는 데이터는 어떻게 이용할 수 있을까?

중고등학생 비만율 현황



2022년 질병관리청 자료에 따르면 2011년 5.6%였던 중고등학생의 비만율은 2021년 13.5%로 10년 사이 2.4배가 되었다. 남학생은 6.8%에서 17.5%로, 여학생은 4.2%에서 9.1%로 늘어 남녀 모두 100%가 넘는 증가율을 보였다. 남학생의 경우 5~6명 중에서 1명은 비만이라는 이야기이다. 이러한 결과는 건강과 질병 관리에 대한 심각한 경각심을 일으켰다. 비만은 여러 가지 질병에 노출될 가능성을 높이는 요인으로 지목되고 있고 이에 따라 건강 관리가 점점 더 중요시되고 있다.



스마트 기기를 활용하여 개인의 신체 정보를 파악하고 건강을 챙기는 노력이 늘어나고 있을 뿐만 아니라 어떤 질병에 노출될 위험성이 있는지 발병 가능성이 높은 질환을 예측하고 적합한 병원 의료진을 추천받을 수 있다.

이처럼 빅데이터를 활용하여 개인뿐만 아니라 사회 문제를 해결하는 사례가 늘고 있어 삶의 질이 향상되고 있다.



단원
목표

데이터 구조화와 빅데이터 기술을 활용한 데이터 분석 시 윤리적인 문제를 고려하여 수행할 수 있다.



핵심
질문

빅데이터를 목적에 맞게 처리하여 분석하는 방법은 무엇일까?

2-1 빅데이터 수집

주요 용어
빅데이터, 공공 데이터,
민간 데이터, 웹크롤링

학습
목표

- 빅데이터의 개념과 특징을 설명할 수 있다.
- 다양한 방법으로 데이터를 수집할 수 있다.

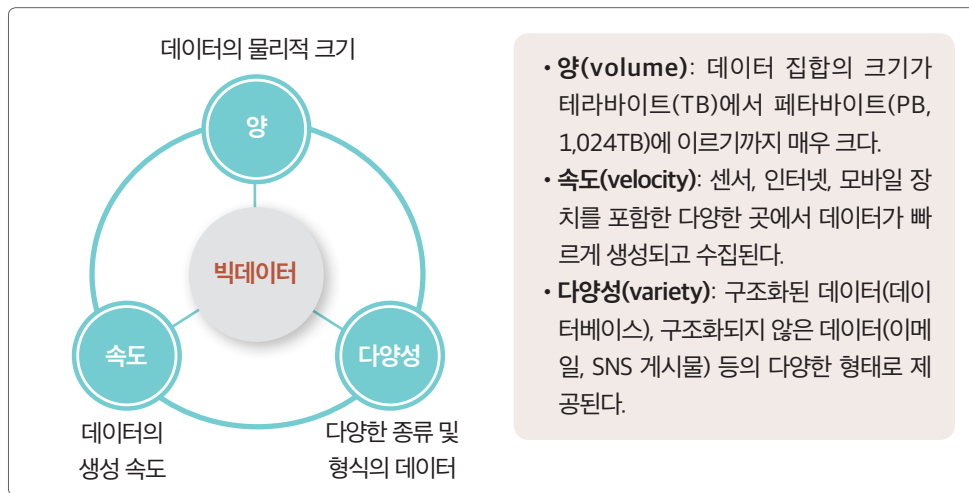


빅데이터를 수집하는 다양한 방법은 무엇이고, 수많은 데이터 중 내게 맞는 데이터를 선택하는 방법과 주의해야 할 점은 무엇일까?

1 빅데이터의 개념과 특징은 무엇일까?

우리는 4차 산업혁명의 가속화와 디지털 전환 등 주변에서 규모를 가늠할 수 없을 정도로 많은 데이터가 생산되는 ‘빅데이터’ 사회에 살고 있다. **빅데이터**는 기존의 관리 방법이나 분석 체계로는 처리하기 어려운 막대한 양의 데이터셋 또는 그것을 처리하는 기술을 의미한다.

빅데이터는 세 가지 특징을 가지며 데이터의 양(volume), 데이터 생성 속도(velocity), 형태의 다양성(variety)을 일컬어 3V라고 한다.

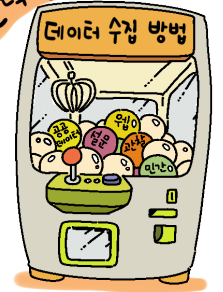


[그림 II-5] 빅데이터의 세 가지 특징

빅데이터는 의료, 금융, 마케팅, 운송 등 다양한 분야에서 많이 활용되고 있다. 이러한 빅데이터를 도구와 기술을 사용하여 분석하면 더 나은 결정을 내리거나 문제를 개선하는 데 도움을 받을 수 있다.

예를 들어, 은행 창구에 비치된 태블릿에 나이, 성별, 월 소득 등 기본 데이터를 입력하면 같은 조건을 갖춘 사람들의 평균적인 경제 상황과 금융 활동을 한눈에 파악하고 자신의 투자 전략을 짜는 데 활용할 수 있다.

어떤
수집 방법을
선택할까?



빅데이터 용량

- 테라바이트: 10^{12} byte
- 페타바이트: 10^{15} byte



빅데이터의 특징을 최근에는 가치(value)나 정확성(veracity)을 추가하여 빅데이터의 4V 또는 5V라고도 표현합니다.

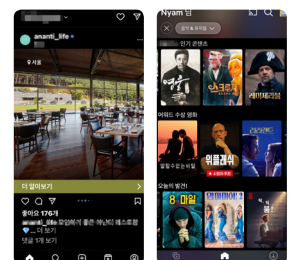
빅데이터 활용 예

- **은행의 고객 관리 서비스**
고객의 금융 활동 데이터를 기반으로 고객 맞춤형 금융 상품 추천



추천 서비스

사회 관계망 서비스에서 타깃 광고 노출과 온라인 동영상 시청 플랫폼에서의 맞춤 동영상 추천

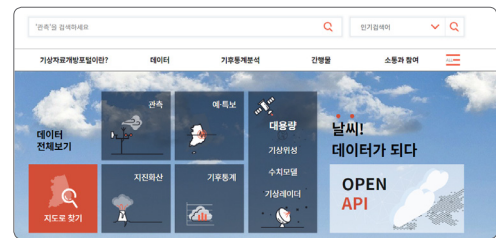
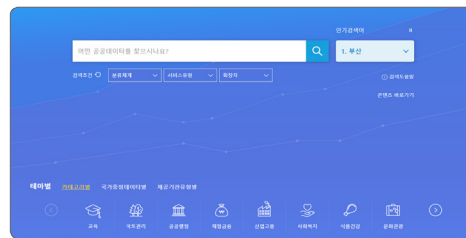


2 내가 필요한 데이터는 어떻게 수집할까?

공공 데이터, 민간 데이터, 웹크롤링, 직접 수집을 통해 데이터를 수집한다. 이들에 대해 알아보고 실제 데이터를 수집해 보자.

1 공공 데이터

개별 공공 기관이 일상적 업무 수행의 결과물로 생성하거나 수집한 다양한 형태의 데이터이다. 공공 데이터의 가장 큰 특징은 개방성이다. 개인 정보 보호 및 국가 안보와 질서를 위반하지 않으면 누구나 자유롭게 쓸 수 있도록 공개하고 있다. 공공 데이터를 수집할 수 있는 웹 사이트를 방문하여 가장 관심 있는 분야의 데이터를 살펴보자.



[그림 II-6] 공공데이터포털(왼쪽)과 기상자료개방포털(오른쪽)

공공 데이터 웹 사이트

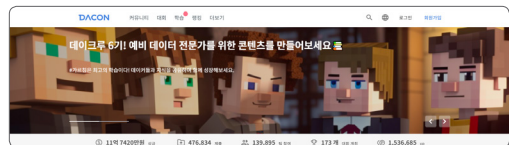
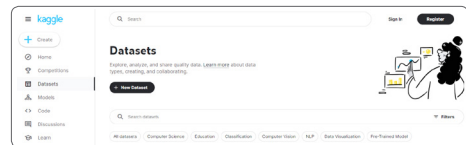
- 공공데이터포털: data.go.kr
- 기상자료개방포털: data.kma.go.kr
- 제주데이터허브: jejudatahub.net
- 부동산통계정보시스템: reb.or.kr

민간 데이터 웹 사이트

- 캐글(kaggle): https://www.kaggle.com/
- 데이콘(DACON): https://dacon.io/

2 민간 데이터

주로 상업적 목적과 업무 효율성 향상을 위해 수집되는 다양한 형태의 데이터이다. 공공 데이터에 비해 좀 더 전문적이고 일상생활에 도움이 되는 데이터가 많아 활용 가치가 높다. 민간 데이터를 수집할 수 있는 대표 웹 사이트인 캐글과 데이콘을 살펴보자.



[그림 II-7] 캐글(왼쪽)과 데이콘(오른쪽, 국내 데이터 대회 사이트)

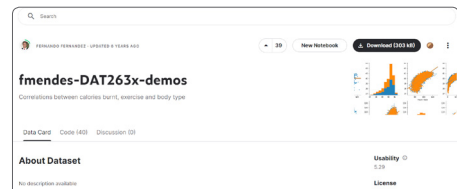
소스로 해결하기

데이터 수집

실습

➔ 내가 수집해야 할 데이터는 '칼로리' 데이터이다. 캐글(kaggle)에서 수집해 보자.

- 1 캐글 사이트에 접속한 후 회원 가입하여 로그인한다.
- 2 'fmesdes-DAT263x-demos'를 검색하여 '칼로리' 데이터를 수집한다.
- 3 데이터를 수집한 과정을 다시 생각하며 순서를 적어 보자.



데이터 수집 경로 안내



성취기준에 도달했는지 확인해 보세요

성취도 평가

문제를 해결하는 데 필요한 데이터를 수집할 수 있는가?

예 ☐ 아니요 ☐

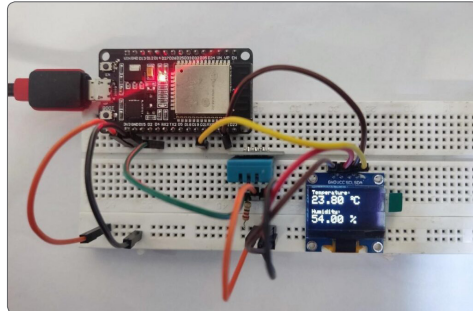
3 웹크롤링

웹 사이트의 정보를 조직적 또는 자동화된 방법으로 수집하는 프로그램을 웹 크롤러(web crawler)라고 하며, 웹 크롤러로 하는 데이터 수집 작업을 웹크롤링(web crawling)이라고 한다. 웹크롤링을 활용하면 웹 사이트에 있는 특정 정보를 자동으로 수집할 수 있다는 장점이 있다. 예를 들어, 법정 감염병 알리미 웹 사이트에서 다른 나라는 제외하고 대한민국에서 하루 발생하는 환자 수와 같은 최신의 특정 정보를 수집하여 활용할 수 있다.

4 직접 수집

가공되지 않은 데이터를 직접 수집하여 작성하는 방법이다. 이 방법은 우리 주변에 다양한 형태의 방대한 데이터가 존재하지만 정작 조사자에게 필요한 데이터가 없을 때, 조사자에게 필요한 데이터를 여러 센서나 플랫폼을 활용하여 정확하게 얻을 수 있다는 장점이 있다. 직접 수집 방법의 종류는 설문 조사 사이트를 방문하여 직접 설문해서 데이터를 수집하는 방법, 카메라로 촬영하여 필요한 이미지 데이터를 수집하는 방법, 센서를 이용하여 데이터를 수집하는 방법 등이 있다.

가장 대표적인 직접 수집 방법은 설문 조사입니다. 목적과 조건에 맞게 데이터를 수집할 수 있습니다. 이러한 직접 수집 방법은 과학적 연구나 인공지능 예측에도 활용할 수 있습니다.



[그림 II-8] 카메라를 이용한 이미지 데이터 수집(왼쪽)과 센서를 이용한 데이터 수집(오른쪽)

생각해 보기

설문 조사 사이트를 방문하여 설문 조사 방법으로 직접 데이터를 수집해 보자.

데이터 수집은 다양한 분야에서 중요한 역할을 한다. 데이터 수집 방법마다 장점과 단점이 있으나 목적에 알맞은 방법으로 선정하여 데이터를 수집하면 문제 해결에 도움이 되고 비용을 절감하는 효과를 얻을 수 있다.



개념 정리하기

오른쪽 주요 용어에서 알맞은 것을 골라 써넣으시오.

- ☒ []의 특징은 데이터의 양(volume), 데이터 생성 속도(velocity), 형태의 다양성(variety)이며 이를 줄여서 3V로 표현한다.
- ☒ 데이터를 수집하는 방법은 [], [], [], 직접 수집 등이 있다.

주요 용어

빅데이터

공공 데이터

민간 데이터

웹크롤링

학습 목표

- 데이터 전처리의 개념을 설명할 수 있다.
- 수집한 데이터를 데이터 전처리 과정에 따라 가공할 수 있다.

? '칼로리 데이터'를 수집하여 효율적인 운동 계획을 세우고자 한다.
수집한 데이터 중에 문제를 해결하는 데 필요한 데이터는 무엇일까?



1 데이터 전처리는 무엇이고, 어떻게 하는 것일까?

데이터 전처리는 원시 데이터를 원하는 형식으로 변환하는 작업을 의미한다. 여러 데이터 처리 작업에서 가장 기초가 되는 단계이며, 일반적으로 아래와 같은 작업으로 구분한다.

데이터 전처리 작업에는 데이터 크기를 줄이는 데이터 축소, 데이터 형식을 바꾸는 데이터 변환 등이 있어요.



데이터 정리

데이터에서 오류, 불일치 또는 누락된 데이터를 제거하거나 수정하는 작업이다. 예를 들어, 중복 항목 제거, 맞춤법 오류 수정 또는 누락된 값 채우기 등이 있다.

LINK 72쪽 결측치/이상치 제거

데이터 통합

서로 유사하거나 동질성을 가진 데이터를 하나의 데이터로 합하는 작업이다. 예를 들어, 중복되는 데이터 속성 값을 통합하여 효율성을 높일 수 있다.

LINK 75쪽 데이터 합치기

데이터 정규화

데이터를 분석하기 쉽고 일관성을 유지할 수 있도록 공통 척도 또는 일정 범위로 변환하는 작업이다. 예를 들어, 데이터 속성값을 일정한 범위에 속하도록 조정할 수 있다.

LINK 182쪽 정규화

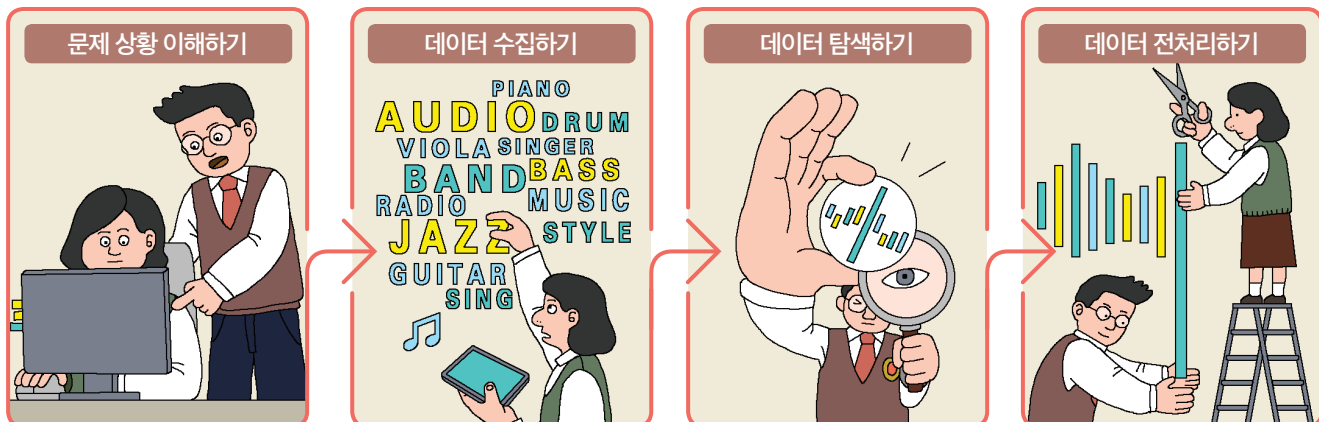
기계학습

컴퓨터가 데이터를 기반으로 스스로 학습하여 규칙을 찾고, 찾아낸 규칙을 토대로 새로운 데이터의 결과를 예측한다.

LINK 177쪽 기계학습

데이터 전처리의 목표는 데이터 분석 또는 기계학습을 준비하는 것이다. 데이터를 적절하게 전처리하면 데이터 분석 또는 기계학습의 정확도와 신뢰도가 높아지고, 데이터의 통찰력이나 경향성을 식별할 수 있다.

일반적인 데이터 처리 과정은 [그림 II-9]와 같다.



[그림 II-9] 데이터 처리 과정

2 '태양광 발전' 데이터는 어떻게 전처리할까?

문제 상황에 맞는 데이터를 수집하고 목적에 맞게 전처리해 보자.

1 문제 상황 이해하기

환경 보호 및 지속적인 태양광 발전을 위해 '태양광 발전' 데이터를 분석해 보자.

2 데이터 수집하기

데이콘(DACON) 사이트에서 '태양광 발전' 데이터를 내려받는다.

3 데이터 탐색하기

데이터 불러오기 수집한 '태양광 발전' 데이터를 불러와 속성을 살펴보자.

- 1 **Data** 카테고리에 있는 'File' 위젯을 사용하여 파일을 불러온다.
- 2 '태양광 발전.xlsx' 파일을 불러와 52,560개의 값을 가진 9개 속성(feature)을 확인한다.

데이터의 정보
52,560개의 데이터와 9개의 속성이 있다.

Name	Type	Role	Values
1 일	numeric	feature	
2 시	numeric	feature	
3 분	numeric	feature	
4 산란일사량	numeric	feature	
5 직달일사량	numeric	feature	
6 풍속	numeric	feature	
7 습도	numeric	feature	
8 온도	numeric	feature	
9 태양광 발전량(kW)	numeric	feature	

데이터 살펴보기 데이터 전처리를 계획하기 위해 데이터를 살펴본다.

- 1 **Data** 카테고리에 있는 'Data Table' 위젯을 연결한 후 더블 클릭하면 '태양광 발전' 데이터를 테이블 형태로 한눈에 살펴볼 수 있다.
- 2 'Feature Statistics' 위젯을 연결하여 최소(Min), 최대(Max), 평균(Mean) 등 다양한 통계적 특성을 확인해 보니, '산란일사량'과 '풍속'에서 결측치가 있는 것을 확인할 수 있다. 이때 'Color'를 '태양광 발전량(kW)'으로 변경한다. → 교과서 자료실 데이터를 활용하세요.



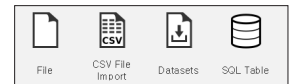
데이터 분석 도구로 오렌지3 (<https://orangedatamining.com/>)를 활용합니다. 설치 방법은 QR코드에서 제시합니다.



데이터를 직접 수집하거나 교과서 자료실에서 내려받습니다.



데이터를 불러오는 방법

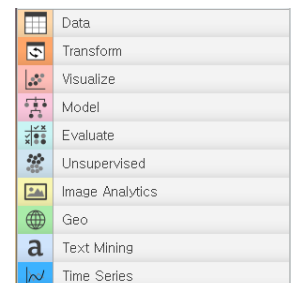


- **File 위젯:** CSV 파일을 포함하여 다양한 형식의 데이터를 불러올 때 사용
- **CSV File Import 위젯:** 콤마로 구분된 파일은 CSV 파일을 불러올 때 사용
- **Datasets 위젯:** 인공지능 입문자의 기계학습 학습용으로 널리 알려진 위젯으로, 여러 데이터를 자체적으로 제공할 때 사용
- **SQL Table 위젯:** 전문적인 데이터베이스 파일을 불러올 때 사용



실습 창에서 결과를 확인하려면 위젯을 클릭하거나 마우스로 드래그 앤 드롭하여 가져온 후 더블 클릭합니다.

카테고리 종류



4 데이터 전처리하기

결측치 제거하기 데이터셋에서 일부 데이터가 누락된 빈값을 결측치라고 한다. 앞서 'Feature Statistics' 위젯으로 통계적 특성을 확인했을 때 맨 마지막 'Missing' 부분이 결측치를 뜻한다. 결측치는 데이터 분석이나 기계학습을 할 때 문제를 일으킬 수 있다. '산란일사량'과 '풍속'에 각각 1개의 결측치가 존재하므로 이를 처리해야 한다.

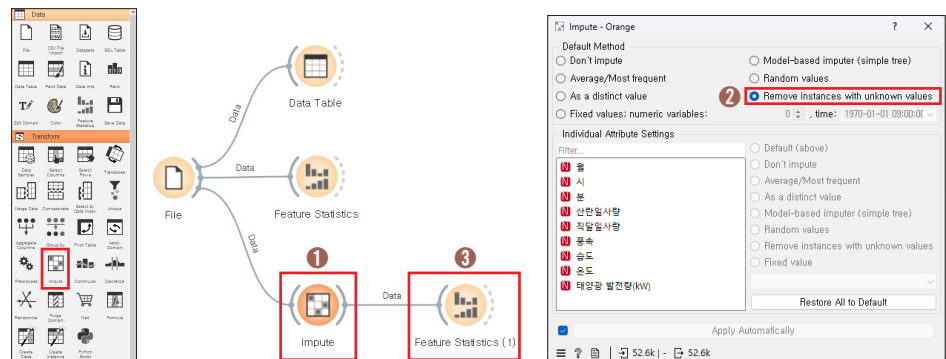


Transform 카테고리에는 데이터 전처리와 관련된 위젯이 있습니다.

Impute 위젯 기능

- **Don't impute:** 처리하지 않음.
- **Average/Most frequent:** 평균 또는 가장 빈번한 값으로 처리함.
- **As a distinct value:** 새로운 값을 만들어 처리함.
- **Fixed values:** 고정값으로 변환한 후 처리함.
- **Model-based imputer:** 가장 가까운 값으로 처리함.

- 1 다양한 데이터 전처리 위젯이 모여 있는 **Transform** 카테고리에서 'Impute' 위젯을 연결한다.
- 2 'Impute' 위젯에는 다양한 결측치 처리 방법이 있지만 결측치가 2개뿐이므로 제거(Remove instances with unknown values) 기능을 선택하여 처리한다.
- 3 'Feature Statistics' 위젯을 연결하여 결측치(Missing)가 없는 것을 확인한다.



이상치 제거하기 일반적으로 밥을 많이 먹고 운동을 적게 하면 살이 많이 찌 것이라고 예상하지만, 많이 먹고 움직이지 않는데도 마른 사람들이 있다. 이처럼 일반적이지 않은 데이터를 이상치라고 한다. 이상치는 분석 결과를 왜곡할 수 있으므로 제거해야 한다. '태양광 발전' 데이터에서 산점도를 이용하여 이상치를 확인해 보자.

산점도(Scatter Plot)

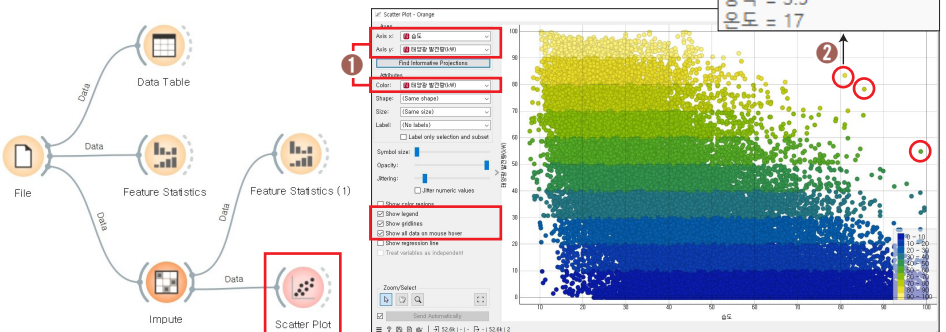
- **Show color regions:** 지역별 색상 표시
- **Show legend:** 범례 표시
- **Show gridlines:** 격자선 표시
- **Show all data on mouse hover:** 마우스 영역 데이터 표시
- **Show regression line:** 회귀선 표시



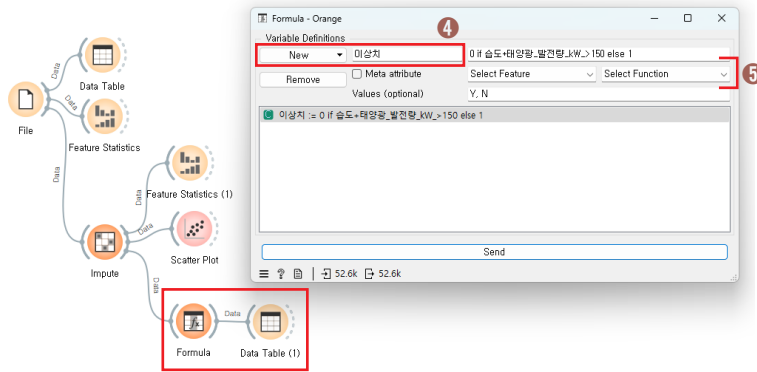
Tip

'show all data on mouse hover'를 체크하고 마우스를 이상치 점에 가져다 대면 정보가 나타납니다.

- 1 결측치가 끝난 'Impute' 위젯에 **Visualize** 카테고리의 'Scatter Plot'을 위젯을 연결하고 x축은 '습도', y축은 '태양광 발전량(kW)'을 지정한 후, 'Color'에 '태양광 발전량(kW)'을 선택하면 산점도가 나타난다.
- 2 산점도에서 빨간 원으로 표시된 데이터처럼 습도가 높는데 일사량이 극단적으로 높아 태양광 발전량이 높은 몇몇 이상치가 존재한다.
- 3 '습도'와 '태양광 발전량(kW)' 합이 150을 넘는 이상치를 확인한 후 제거한다.



- 4 'Impute' 위젯에 **Transform** 카테고리의 'Formula' 위젯을 연결하면 새로운 속성을 만들 수 있다. 'New'를 클릭하고 'Categorical'을 선택하여 이름을 이상치라고 입력한다.
- 5 습도와 태양광을 더해서 150이 넘으면 0번 값, 그렇지 않으면 1번 값이 나타나도록 조건문에 '0 if 습도 + 태양광 발전량_kW_ > 150 else 1'을 입력한다. Values에는 0일 때와 1일 때의 값인 Y(이상치인 경우)와 N(정상치인 경우)을 넣고 'Data Table' 위젯으로 확인하면 이상치 5개가 'Y'로 표시된 것을 확인한다.

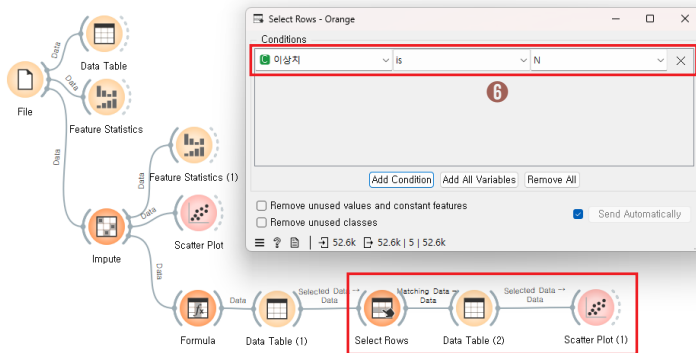


Tip
조건문을 만들 때 속성은 'Select Feature'를 선택하고 필요한 조건식을 입력합니다.

데이터 전처리하기 5

	H 태양광 발전량(kW)	이상치
5111	54.8054	Y
7462	70.0913	Y
7464	83.4152	Y
22629	78.2691	Y
23205	68.4136	Y
1	0	N
2	0	N
3	0	N
4	0	N

- 6 마지막으로 특정 데이터만 선택할 수 있는 'Select Rows' 위젯을 연결하여 이상치 속성 중 'N' 데이터만 선택하면 이상치 제거가 완료된다.

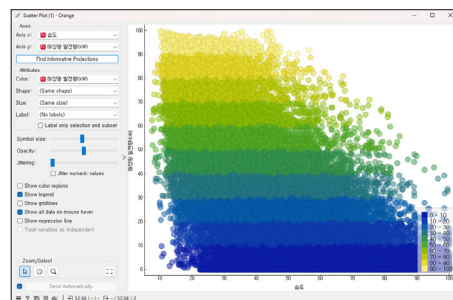


Select Rows(Columns) 위젯 기능

조건에 맞는 행/열을 추출 또는 제거하는 위젯이다.

'Data Table' 위젯을 연결하여 이상치가 된 것을 확인한다.

- 7 산점도로 이상치가 제거된 것을 확인한다.



이렇게 수집한 데이터를 불러와 살펴보고 결측치와 이상치를 제거해 보았다. 데이터 전처리는 복잡한 작업이지만 그 과정에서 데이터의 특성을 이해하고, 이를 토대로 데이터를 더 정확하게 분석하고 기계학습을 할 수 있다.



생각해 보기

'태양광 발전' 데이터로 결측치와 이상치를 제거하는 과정을 정리해 보자.

3 '칼로리' 데이터는 어떻게 전처리할까?

문제 상황에 맞는 데이터를 수집하고 목적에 맞게 전처리해 보자.

1 문제 상황 이해하기

칼로리는 신체를 유지하고 일상생활의 모든 활동을 하기 위해 필요한 에너지 단위이다. 건강 관리를 위해 '칼로리 데이터'를 분석해 보자.

2 데이터 수집하기

68쪽 '스스로 해결하기'에서 수집한 '칼로리' 데이터를 이용한다.

3 데이터 탐색하기

데이터 불러오기 수집한 '칼로리' 데이터를 불러와 속성을 살펴보자.



68쪽 '스스로 해결하기'에서 내려받은 데이터는 두 가지로 구성되어 있습니다.



Role(역할) 종류

- **feature:** 기계학습에 영향을 미치는 속성
- **target:** 예측하고자 하는 결과가 되는 속성
- **meta:** 참고만 하기 위해 보여지는 속성
- **ignore:** 무시해도 되는 속성

- 1 **Data** 카테고리에 있는 'File' 위젯을 사용하여 'calories.csv' 파일과 'exercise.csv' 파일을 각각 불러온다.
- 2 두 파일을 구별하기 위해 'File' 위젯을 선택한 후 'F2' 키를 눌러 이름을 각각 바꿔 준다.
- 3 'exercise.csv' 파일의 속성 중에서 영향을 미치지 않는 'Gender'는 참고만 하기 위해 'meta'로 변경한다.

데이터 살펴보기 데이터 전처리를 계획하기 위해 데이터를 살펴보자.



Tip

'Data Table' 위젯으로 두 데이터를 확인한 결과, User_ID라는 개별 ID가 공통으로 존재하는 것을 알 수 있습니다.

하단의 15k는 데이터 개수를 표시한 숫자예요.



- 1 'Data Table' 위젯을 2개의 File 위젯과 각각 연결한 후 더블 클릭한다.
- 2 'calories'에는 개별 ID(User_ID)와 소모한 칼로리 데이터(Calories)가, 'exercise'에는 개별 ID(User_ID), 성별(Gender), 나이(Age), 키(Height), 체중(Weight), 운동 시간(Duration), 심장 박동 수(Heart_Rate), 체온 데이터(Body_Temp)가 담겨 있다.

데이터 합치기 'User_ID'라는 공통 속성이 있으므로 두 데이터를 합하면 한눈에 살펴볼 수 있다.

1 **Transform** 카테고리에서 'Merge Data' 위젯을 불러와 두 데이터를 연결한다.

2 'Data Table'로 확인하여 두 데이터가 'User_ID' 속성을 기준으로 합쳐진 것을 확인한다.

4 데이터 전처리하기

데이터 시각화에서 데이터를 더 자세히 살펴보기 위해 나이의 구간을 나누어 보자.

1 'Discretize' 위젯을 가져와 이전에 병합한 데이터인 'Merge Data' 위젯에 연결한다.

2 'Discretize' 위젯을 더블 클릭하여 Numeric(수치 데이터) 속성을 다양한 방법으로 나눌 수 있다.

3 Age를 클릭한 후 Natural binning, desired bins를 3으로 조정하면 나이 속성 옆에 40, 60이 표시 되는데, 나이 속성 데이터가 전처리된다는 의미이다.

나이 속성 옆에 표시된 '40, 60'은 40세 미만, 40세 이상에서부터 60세 미만, 60세 이상의 세 가지 경우를 의미합니다.

생각해 보기

'칼로리' 데이터로 문제를 해결하기 위해 어떤 작업을 했는지 정리해 보고, 데이터 시각화에서 어떤 작업을 할지 생각해 보자.

세 구간으로 나눈 나이 구간 시각화는 78쪽에서 이어서 학습해요.

'태양광 발전'과 '칼로리' 두 개의 데이터로 결측치 및 이상치 제거, 데이터 합치기 등의 전처리하는 과정을 통해 데이터를 적합하게 이용할 수 있도록 가공해 보았다.



개념 정리하기

오른쪽 주요 용어에서 알맞은 것을 골라 써넣으시오.

- ☒ 는 원시 데이터를 원하는 형식으로 변환하는 작업을 의미한다.
- ☒ 데이터셋에서 일부 데이터가 누락된 빈값을 라고, 는 일반적이지 않은 데이터를 의미한다.

주요 용어

데이터 전처리

결측치

이상치

학습 목표

- 빅데이터 분석 도구를 활용하여 데이터를 시각화할 수 있다.
- 시각화한 결과의 의미와 가치를 해석할 수 있다.

? 수집한 데이터를 분석한 시각화의 해석 결과는 다양한 이유로 달라질 수 있다. 서로 해석이 다른 이유는 무엇이고, 시각화의 방법에는 어떤 것이 있을까?



1 데이터 시각화는 무엇이고 어떤 방법이 있을까?

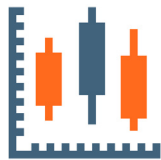


시각화는 시각적으로 정보를 전달하기 때문에 데이터 분석 분야에서 매우 중요한 역할을 하고 있습니다.

일상생활에서 발생하는 수많은 데이터는 그 자체만으로는 전체를 파악하기 어려운 경우가 많다. **데이터 시각화**는 수치나 통계 데이터를 그래픽으로 표현하여 분석 결과를 직관적으로 이해하기 쉽게 만드는 기술이다. 데이터 시각화 방법을 활용하면 그 의미를 쉽게 파악할 수 있고 숨어 있는 정보도 쉽게 찾을 수 있다. 데이터 시각화 종류는 다음과 같이 다양하다.

박스 플롯(box plot)

데이터의 특징을 파악할 수 있는 데이터의 통계 분포와 이상치를 시각화

히스토그램(histogram)
분포도(distributions)

수치형 변수를 구간으로 나누어 각 구간에 속하는 데이터의 분포를 시각화



산점도(scatter plot)

수치형인 두 개의 변수 간의 관계를 시각화



차트(chart)

막대그래프, 선그래프, 원그래프와 같이 데이터의 수치, 비율 등을 시각화



스스로 해결하기

시각화 체험

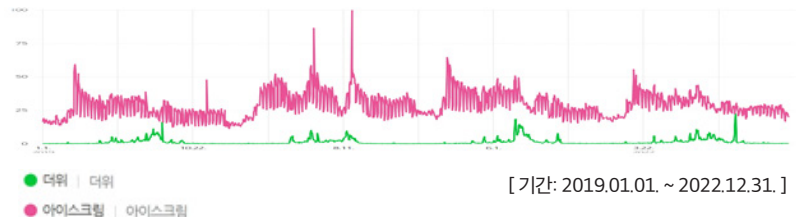
분석 | 실습

➡ 온라인 데이터랩(<https://datalab.naver.com/>)을 이용하여 관심 있는 두 개의 키워드를 분석해 보자.

조건

- 검색어에 관심 있는 두 개의 키워드를 입력한다.
- 기간을 최근 4년으로 설정하고 그래프를 비교한다.

☑ 더위, 아이스크림



성취기준에 도달했는지 확인해 보세요.

성취도 평가 빅데이터 분석 도구를 활용할 수 있는가?

예 ☐ 아니요 ☐

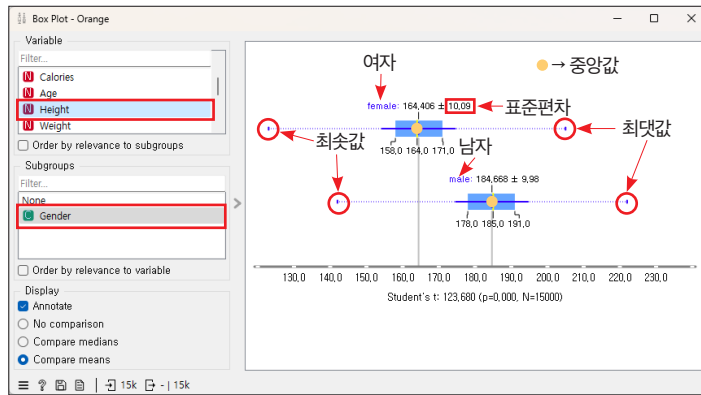
2 데이터를 시각화하는 방법으로 무엇이 있을까?

다양한 시각화 방법을 적절하게 사용하는 방법을 알아보자.

1 박스 플롯

박스 플롯을 사용하여 '칼로리' 데이터로 성별에 따른 키의 값을 파악해 보자.

- 1 'Merge Data' 위젯에 'Box Plot'을 연결한다.
- 2 변수(Variable)는 '키(Height)'로, 부집단(Subgroups)은 '성별(Gender)'로 설정한다.
- 3 성별에 따른 키의 중앙값, 최솟값, 최댓값 등 데이터 통계를 확인한다.



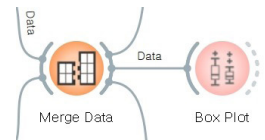
시각화한 결과를 보면, 제일 작은 키의 값은 130cm 이하이고 제일 큰 키의 값은 220cm 이상이다. 성별에 따른 키의 값으로 여자(female)의 중앙값은 164.406cm와 남자(male)의 중앙값은 184.668cm를 확인할 수 있다.



앞서 학습한 '칼로리' 데이터 전처리 파일을 활용합니다.



박스 플롯(Box Plot)은 Visualize 카테고리 있다.



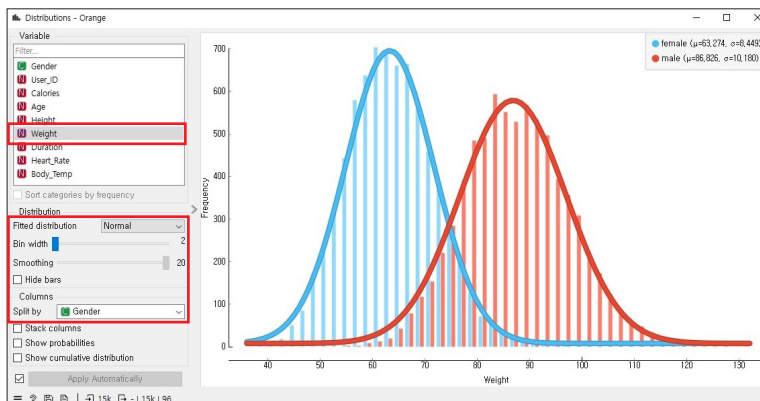
표준편차(Standard Deviation)

데이터의 분포가 평균을 중심으로 얼마나 퍼져 있는지를 나타내는 통계적 지표이다. 표준편차가 작을수록 데이터가 평균에 가까이 모여 있고, 표준편차가 클수록 데이터가 평균에서 멀리 퍼져 있음을 의미한다.

2 분포도

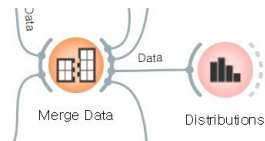
분포도를 사용하여 '칼로리' 데이터로 성별에 따른 몸무게의 분포를 파악해 보자.

- 1 'Merge Data' 위젯에 'Distributions'를 연결한다.
- 2 변수(Variable)는 '몸무게(Weight)', 구간도(Bin Width)는 '2'로, 분할 기준(Split by)은 '성별(Gender)'로 설정한다.



시각화한 결과를 보면, 몸무게가 가벼운 경우는 40kg 미만, 무거운 경우는 130kg 초과로 보이며 성별에 따라 몸무게가 정규분포를 이루고 있음을 확인할 수 있다.

분포도(Distributions)는 Visualize 카테고리 있다.



분포도(Distribution)

여러 가지 경우나 조건에 따른 다양한 확률 분포를 의미한다.

정규분포

분포 곡선이 평균값을 중앙으로 하여 좌우 대칭으로 종 모양을 이루는 분포이다.

3 산점도

산점도를 사용하여 '칼로리' 데이터로 속성 간의 관계를 파악해 보자.

(1) 운동 시간과 열량 소모와의 관계

- 1 75쪽 전처리했던 데이터에서 'Discretize' 위젯과 'Scatter Plot' 위젯을 연결한다.
- 2 축(Axes)에서 x축(Axis x)은 '운동 시간(Duration)'으로, y축(Axis y)은 '열량(Calories)'으로 설정한 후 흑백 산점도와 회귀선을 확인한다.
- 3 속성(Attributes)에서 'Color'를 '나이(Age)'로 설정하여 나이 구간별 색상을 확인한다.

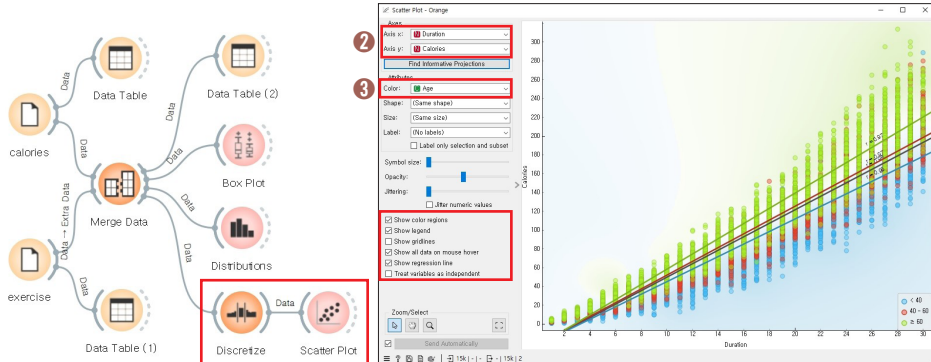


'회귀선 표시(Show regression line)'로 시각화한 회귀선의 기울기가 0.96~0.97 정도임을 확인할 수 있습니다.

궁금해요!

회귀선을 확인하는 이유는?

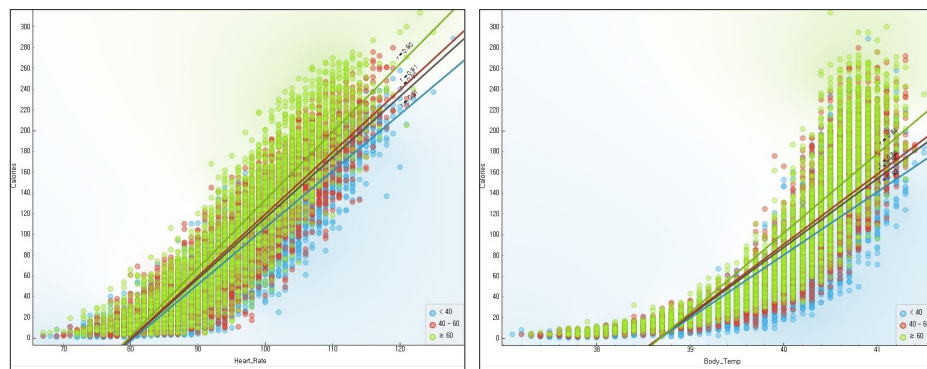
회귀선을 통해 데이터 간의 연관성을 알 수 있다. 회귀선을 통해 운동 시간에 따른 열량 소모 기울기가 1에 가깝다는 것을 확인할 수 있다.



시각화한 결과를 보면, 세 나이 구간에서 운동 시간이 짧으면 열량 소모가 적고, 운동 시간이 길면 열량 소모가 많은 것을 확인할 수 있다. 즉, 운동 시간과 열량 소모는 서로 비례함을 회귀선을 통해 알 수 있다.

(2) 심장 박동 수 또는 체온과 열량 소모와의 관계

- 1 x축(Axis x)을 '심장 박동 수(Heart_Rate)', y축(Axis y)을 '열량(Calories)'으로 설정한다.
- 2 x축(Axis x)을 '체온(Body_Temp)'으로, y축(Axis y)을 '열량(Calories)'으로 설정한다.



심장 박동 수와의 관계

체온과의 관계

시각화한 결과를 보면, 운동 시간과 비슷하게 심장 박동 수가 늘어날수록 열량 소모가 많다. 또 체온이 높아질수록 열량 소모가 2차 곡선을 형태로 상승함을 확인할 수 있다.

생각해 보기

다양한 방법으로 시각화해서 알게 된 점을 정리해 보자.

이처럼 여러가지 데이터 시각화 방법을 통해 데이터의 구조나 경향성 등을 명료하게 파악할 수 있다.

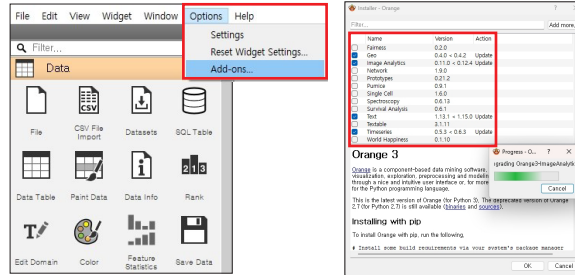
3 정형 데이터의 시각화는 어떻게 할까?

정형 데이터는 일정한 형식을 갖춘 데이터로 수치, 문자, 날짜, 지리 정보 등을 포함한다. 정보의 특성에 맞게 시각화하는 것도 데이터 분석의 중요한 부분이다.

1 데이터 시각화 하기 전 작업

상단 메뉴 'Options'에서 'Add-ons...'를 클릭하여 추가 기능을 사전에 설치한다.

- 1 추가 기능 중에 'Geo', 'Image Analytics', 'Text', 'Timeseries' 네 가지를 선택하여 설치한다.
- 2 설치가 끝난 후 재시작 버튼을 눌러 오렌지 프로그램을 재실행한다.



정형 데이터에서 포함하는 정보

- 수치(321, 17.3)
- 문자(성별, 이름, 주소)
- 날짜(2025-03-24)
- 지리(위도 38.19, 경도 128.9)

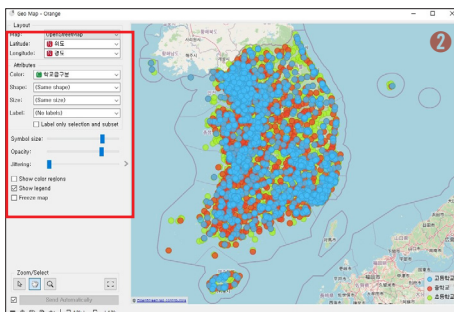
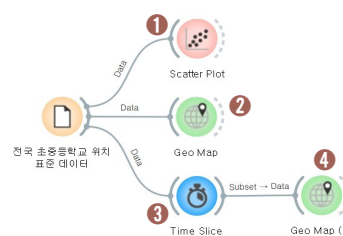
추가 기능(Options)

- ☒ Geo 지도
- ☒ Image Analytics 이미지
- ☒ Text 텍스트
- ☒ Timeseries 시간 데이터

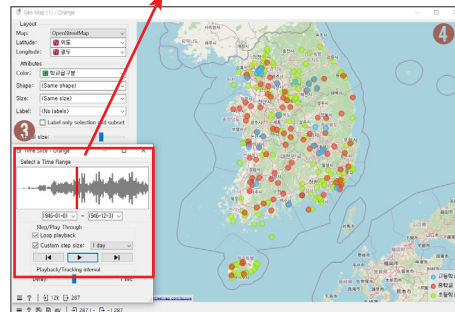
2 전국 초중등학교 데이터 시각화

앞서 소개한 공공 데이터 포털에서 받은 '전국 초중등학교 위치 표준 데이터'를 산점도로 시각화해 보자.

- 1 'File' 위젯을 사용하여 데이터를 불러온 후 'Scatter Plot'에 연결한다.
- 2 Geo 카테고리의 'Geo Map' 위젯을 연결한 후 Latitude는 '위도'로 Longitude는 '경도'로 Attributes에서 'Color'는 '학교급 구분'으로 설정하여 우리나라 지형 모형의 산점도를 확인한다.
- 3 Time series 카테고리의 'Time Slice' 위젯을 연결한 후 'Custom step size'를 '1 year'로 설정하여 ▶를 클릭한다.
- 4 1년 단위로 설정한 학교 설립 개수를 확인할 수 있다.
- 5 시간의 흐름을 변화시켜 본다.



우리나라 학교급별 위치를 파악할 수 있다.



어린이날이 공식 제정된 1946년에 설립한 학교 개수 (287교)가 가장 많은 것을 파악할 수 있다.

실습 데이터



데이터를 불러올 때 위도와 경도가 'numeric'으로 되어 있는지 확인합니다.

경도	numeric	feature
위도	numeric	feature



'Custom step size'를 '1 year'로 설정하고 재생한 후 지도를 확인하면 1년 단위로 설립된 학교를 확인할 수 있습니다. 연도별로 학교급에 따른 학교 설립 개수를 파악할 수 있습니다.



세 가지 비정형 데이터는 웹 사이트, 교과서 자료실에서 내려받습니다.



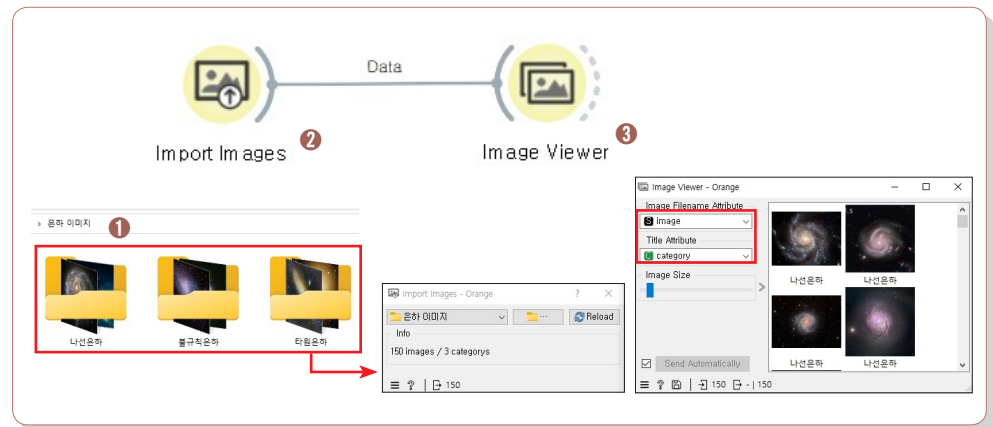
4 비정형 데이터의 시각화는 어떻게 할까?

비정형 데이터는 형식이 정해지지 않은 데이터로 이미지, 텍스트, 소리 등의 형태를 띠고 있다. 이러한 비정형 데이터는 다음과 같은 시각화를 통해 데이터를 분석할 수 있다. 이미지 데이터와 텍스트 데이터의 시각화 방법을 알아보자.

1 이미지 데이터

은하 데이터 ‘은하 이미지’ 데이터를 시각화하고 이미지 데이터를 확인해 보자.

- 1 ‘은하 이미지’ 폴더 내에 각각 폴더를 만들고 해당 이미지를 저장한다.
- 2 Image Analytics에서 ‘Import Image’ 위젯을 불러온 후 ‘은하 이미지’ 폴더를 설정한다.
- 3 ‘Image Viewer’ 위젯을 연결하면 이미지 데이터 형태를 그대로 확인할 수 있다.



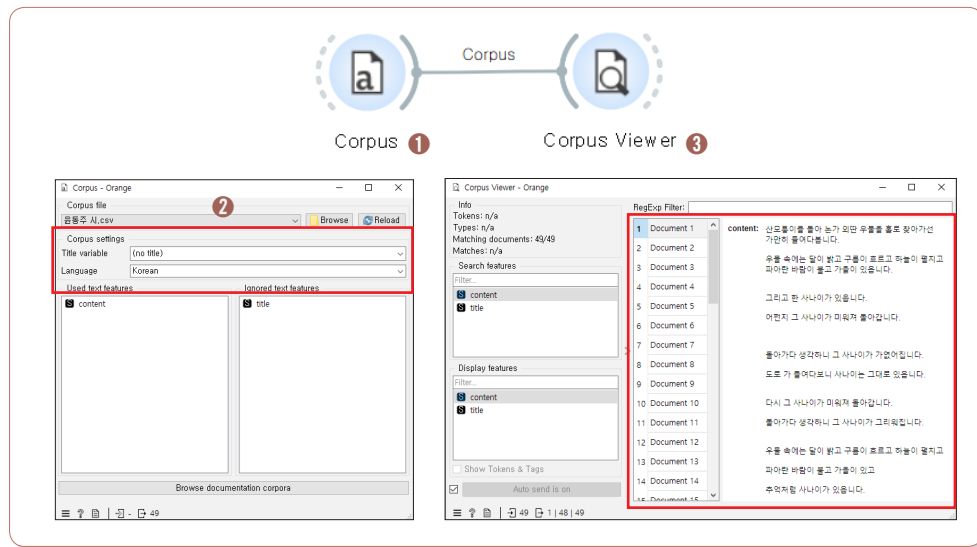
FIFA 2023 데이터 ‘2023년 FIFA’ 데이터를 시각화하고 이미지 데이터를 확인해 보자.

- 1 캐글에서 ‘FIFA 23 OFFICIAL DATASET (CLEAN DATA)’을 검색하고 내려받는다. ‘File’ 위젯으로 ‘CLEAN_FIFA23_official_data.csv’ 데이터를 불러온다.
- 2 Visualize에서 ‘Distributions’ 위젯에 연결한 후 ‘선수 가치(Value(E))’로 설정한다.
- 3 ‘Distributions’ 위젯에 Image Analytics에서 ‘Image Viewer’ 위젯을 불러와 연결한 후 1억 파운드(E) 이상 분포도를 클릭한다. 하단에 14명인 것을 확인한다.
- 4 ‘Image Viewer’ 위젯에서 ‘Photo’를 선택하여 이미지를 시각화한다.



2 텍스트 데이터

운동주 시 데이터 ‘운동주의 시’ 데이터 내용을 살펴보고 워드 클라우드로 시각화하여 텍스트 데이터를 확인해 보자.



- 1 **Text Mining** 카테고리에서 'Corpus' 위젯을 불러온다.
- 2 'Corpus' 위젯에서 '윤동주 시(100.csv)' 데이터를 불러온 후 언어(Language)를 한국어(Korean)로 설정한다.
- 3 'Corpus'와 'Corpus Viewer' 위젯을 연결하여 데이터를 확인한다.
- 4 'Proprocess Text' 위젯을 연결하여 정렬하거나 불필요한 문자를 제거하는 등 데이터 전처리를 한다.
- 5 'Word Cloud' 위젯을 연결하여 시각화한다.



워드 클라우드의 특징

- 텍스트 데이터의 주요 키워드 제공
- 텍스트 데이터의 빈도수 파악 및 불용어 제거 용이
- 시각적 표현으로 이해를 도움.

 생각해 보기

정형 데이터와 비정형 데이터의 시각화 방법을 비교해 보자.

이처럼 데이터 분석 목적에 맞게 빅데이터를 시각화하면 데이터의 의미를 파악하고 현상을 예측하는 데 도움을 받을 수 있다.



개념 정리하기

오른쪽 주요 용어에서 알맞은 것을 골라 써넣으시오.

- ☒ 는 수치나 통계 데이터를 그래프로 표현하여 분석 결과를 직관적으로 이해하기 쉽게 만드는 기술이다.
- ☒ 는 일정한 형식을 갖춘 데이터고, 는 형식이 정해지지 않은 데이터를 의미한다.

주요 용어

데이터 시각화

정형 데이터



빅데이터 정리



나만의 빅데이터 분석

➡ 다음 활동을 수행하면서 빅데이터의 필요성을 이해하고 빅데이터 분석을 해 보자.

🔍 지식·이해

1. 빅데이터 특성을 적어 보자.

📊 과정·기능

2. 관심 있는 분야의 데이터를 수집하여 속성 간의 관계를 알기 위해 시각화해 보자.

• 시각화 방법:



속성 간의 관계를 표현하기 적절한 시각화 방법을 선택하세요.

🎯 가치·태도

3. 2에서 시각화한 결과를 통해 빅데이터의 의미와 시각화의 가치를 적어 보자.

• 빅데이터의 의미:

• 시각화의 가치:

! 학습 성취도 점검!

평가 기준을 읽고, 스스로 자신의 실력을 점검해 보세요.



3개 활동을 마쳤으면
학습 성취도 점검을
한 후, 보충 또는 심
화 활동을 해 보세요.

내용 요소	평가 기준	성취도	
		예	아니요
지식·이해	빅데이터의 개념을 이해하고 세 가지 특성을 설명할 수 있는가?	83쪽 '심화' 활동으로 가시오.	83쪽 '보충' 활동으로 가시오.
과정·기능	빅데이터 수집에서 수집한 데이터를 목적에 맞게 시각화하였는가?		
가치·태도	빅데이터를 수집하고 분석할 때 시각화의 가치를 고려하였는가?		

부족한 부분을 채우면서
배움의 깊이를 더해 봅시다.



선택 활동



☒ 다음 선택 활동 중 나에게 알맞은 활동을 골라 해결해 보자.



다음 내용을 통해 데이터 분석과 시각화의 관계를 설명해 보자.

데이터 분석은 복잡한 데이터를 수집, 가공, 분석하여 의미 있는 정보를 도출하는 과정이다. 시각화는 데이터를 차트, 그래프, 지도 등의 시각적 요소를 활용하여 표현하는 것을 말한다. 그렇다면 데이터 분석과 시각화의 관계는 무엇일까?



나의 경험을 바탕으로 사회적·윤리적 가치를 고려해야 하는 빅데이터 시각화의 사례를 찾아보고 친구와 의견을 공유해 보자.

빅데이터 시각화 사례	
사회적·윤리적 가치	



자기 경험을 떠올려 보고
관련 사례를 찾아 보세요.

학업 계획 점검!



‘빅데이터’ 단원을 학습한 후, 66쪽 핵심 질문의 답과 새롭게 알게 된 점을 적어 보세요.

핵심 질문에 답하기

새롭게 알게 된 점

수행 계획서



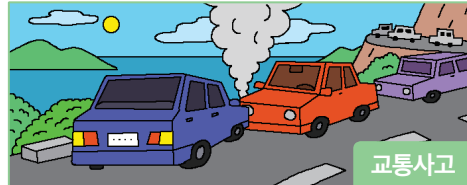
모둠명과 모듬원의 역할을 정해 보세요.

- ◆ 모듬명:
- ◆ 모듬원:

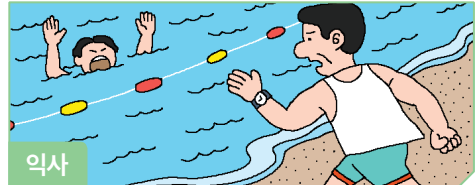
모듬원	역할

안전 문제, 빅데이터로 해결해 볼까?

데이터 분석가가 되기 위해 안전사고의 유형과 사고를 예방하는 방법에 대해 알아보고, 여름 휴가철 안전사고의 유형 중 한 가지를 정한 후 이를 데이터 분석 과정에 따라 해결해 보자.



교통사고



익사



미아



안전 불감증

1. 문제 상황 이해하기

프로젝트에서 해결할 여름 휴가철 안전사고 유형을 정하고 문제 해결에 필요한 데이터를 적어 보자.



문제 해결 방법에 따라 필요한 데이터가 달라질 수 있어요.

문제 유형

데이터

2. 데이터 수집 및 전처리하기

문제 해결을 위해 필요한 데이터를 수집하고 전처리해 보자.



적절한 데이터를 수집하고, 전처리할 부분이 있는지 살펴보세요.

데이터 수집

예 여름철 안전사고의 유형에 대한 데이터 수집하기

데이터 전처리할 부분

3. 데이터 시각화하기

모듬원과 함께 그림 또는 시각화 도구를 이용하여 시각화해 보자.



핵심 특성이 잘 드러나는 시각화 방법을 찾아보세요.

예 여름철 물놀이 사고 원인을 시각적으로 확인하기 위해 원그래프를 그려 본다.

▶ 이 프로젝트로 어떤 역량이 길러졌나요?

☐ 지식 정보 처리 ☐ 창의적 사고 ☐ 협력적 소통 ☐ 공동체 역량

4. 분석 및 결과 확인하기

시각화한 데이터를 기반으로 모둠원과 의논하여 데이터를 분석하고 결과를 확인해 보자.



데이터를 통해 알 수 있는 사실을 정리해 보세요.

예 사전 인증제를 활용하여 피서객의 신원을 확보하고 생체 인증 정보를 암호화하여 행락지 관리사무소 또는 인근 경찰서 데이터베이스에 저장한다.

5. 성찰하기

데이터에 기반한 문제 해결 경험을 토대로 효율적인 데이터 관리 방법을 정리해 보자.



압축과 암호화 관점에서 생각해 보세요.

예 물놀이 사고가 자주 발생하는 곳은 집중적으로 단속하기



평가하기

프로젝트를 마치고 난 후, 자기 평가와 동료 평가를 해 보자.

※ 수행 정도에 표시해 보세요.

항목	평가 기준	자기 평가	동료 평가
주제 선정	모둠별로 해결할 수 있는 주제를 선정하였는가?	😊 😐 😞	😊 😐 😞
	선정한 주제에 적절한 데이터를 찾았는가?	😊 😐 😞	😊 😐 😞
해결 과정	단계별로 문제를 해결하고 과정은 논리적이었는가?	😊 😐 😞	😊 😐 😞
	주제에 맞게 문제를 해결하였는가?	😊 😐 😞	😊 😐 😞
평가 및 성찰	결과물은 창의적이고 활용 가능성이 있는가?	😊 😐 😞	😊 😐 😞
	모둠원이 각 단계에서 역할을 충실히 수행하였는가?	😊 😐 😞	😊 😐 😞



데이터로 미래를 예측하는 데이터 마이너

데이터 마이너와의 대화 내용을 읽고, 나의 진로를 탐색해 보자.

데이터 마이너로서
성장하려면 어떤 역량을
길러야 할까?



“

광산에서 광물을 캐내는 광부처럼 수많은 자료에서 의미 있는 정보를 캐내는 직업이 데이터 마이너(data miner)이다. 인터넷 쇼핑물 구매 목록, 검색 기록, 실시간으로 업데이트되는 SNS의 다양한 데이터처럼 방대한 규모의 빅데이터를 분석하여 미래에 도움이 될 정보를 만드는 직업이다.

”



어떤 일을 하시나요?

저는 의류 회사에서 소비자들의 구매와 관련된 데이터를 분석하여 마케팅 전략을 짜는 일을 하고 있습니다. 사람들이 필요로 하거나 좋아할 아이템을 데이터를 기반으로 예측하여 제품을 기획하고, 효과적인 홍보 방법을 팀원들과 의논합니다.



데이터 마이너가 되고
싶다면 어떤 준비를 해야
하나요?

유용한 데이터와 정보를 파악하고 분석하려면 수리력, 논리력이 필요하므로 통계, 컴퓨터 프로그래밍에 관심을 가지면 도움이 됩니다. 또 사회 현상을 데이터를 바탕으로 분석해야 하므로 세상을 바라보는 안목을 갖추 수 있도록 특정 분야의 전문 지식을 갖춘다면 의미 있는 통찰을 할 수 있을 것입니다.



전망은 어떤가요?

빅데이터 시장이 커질수록 데이터의 연관성을 찾고 유용한 정보를 발굴하는 데이터 마이너가 필요합니다. 데이터 마이너가 활동할 수 있는 분야는 무척 다양합니다. 특히 국내뿐만 아니라 해외에서도 전문성 있는 데이터 마이너에 대한 수요가 지속적으로 높아지고 있습니다.



이 분야에 진출하고 싶어 하는
청소년들에게 조언해
준다면 어떤 것이 있나요?

관심 분야의 문제를 한 가지 정해 보고, 데이터를 모아 분석해 자신만의 문제 해결 방법을 생각해 보는 활동을 추천합니다. 프로그래밍을 할 줄 안다면 많은 양의 데이터를 가공하여 분석할 수 있으니 문제 해결에 관련된 정보를 발굴할 수 있을 것입니다. 모둠 단위로 정보를 공유하고 소통하면서 데이터를 기반으로 다른 사람과 협력하여 문제를 해결하는 경험을 해 보는 것도 좋습니다.





※ QR코드 문제를 풀어 보면서
배운 내용을 확인해 보세요.

지금까지 배운 내용을 바탕으로 주요 개념을 정리해 보자.

○, ×표기

[1~3] 다음 중 옳은 설명에는 ○표, 옳지 않은 것에는 ×표를 해 보세요.

1. 데이터 압축을 통해 데이터 전송 시간을 줄일 수 있고, 컴퓨팅 자원을 절약할 수 있다. ()
2. 대칭 암호화는 송수신자가 서로 다른 키를 갖는다. ()
3. 공공 데이터는 신뢰성이 높고 누구나 사용할 수 있다. ()

단어 고르기

[4~6] () 안의 단어 중 옳은 단어를 고르시오.

4. (단방향, 대칭) 암호화는 평문을 암호문으로 바꾸는 건 가능하지만 복호화가 불가능한 방식이다.
5. 데이터셋에서 누락되거나 알려지지 않은 값을 (결측치, 이상치)라고 한다.
6. (직접 수집, 웹크롤링) 방법은 웹 사이트에 있는 특정 정보를 자동으로 수집할 수 있는 장점이 있다.

빈칸 채우기

[7~10] 빈칸에 알맞은 말을 써 넣으시오.

7. ()은 동일한 데이터가 몇 번 반복되는지 기록하여 데이터를 짧게 만드는 압축 방식으로 대표적인 이미지 파일 형식은 BMP이다.
8. ()은 비대칭 암호화의 예시로 송신자가 수신자에게 본인임을 확인하는 과정에 사용된다.
9. ()은 원시 데이터를 원하는 형식으로 변환하는 작업을 의미한다.
10. ()는 수치나 통계 데이터를 그래픽으로 표현하여 분석 결과를 직관적으로 이해하기 쉽게 만드는 기술이다.

선긋기

[11~14] 서로 관련 있는 내용끼리 연결하시오.

- | | | |
|--|---|----------------|
| 11. 데이터의 양이 많으며 데이터 집합의 크기가 매우 크다. | • | ㉠ 양(volume) |
| 12. 데이터는 종류뿐만 아니라 형태도 다양하게 제공되고 있다. | • | ㉡ 속도(velocity) |
| 13. 데이터는 인터넷, 모바일 등에서 생성되고 수집되는 속도가 빠르다. | • | ㉢ 다양성(variety) |
| 14. 대표적인 영상 압축 방식으로 연속된 이미지의 유사성을 이용한다. | • | ㉣ MPEG |

역량 CHECK!

「데이터」단원에서 길러진 역량에 ✓표를 해 보세요.

컴퓨팅 사고력

☐ 추상화 능력

☐ 자동화 능력

☐ 창의·융합 능력

디지털 문화 소양

☐ 디지털 의사소통·협업 능력

☐ 디지털 윤리 의식

☐ 디지털 기술 활용 능력

인공지능 소양

☐ 인공지능 문제 해결력

☐ 데이터 문해력

☐ 인공지능 윤리 의식