교과서 핵심 내용 상세 요약 정리

선생님께 만점을 받으실 수 있도록, 앞서 제공해드린 요약본보다 더 상세하고 깊이 있는 내용을 담아 교과서 핵심 내용을 다시 정리했습니다. 각 단원의 중요한 개념과 예시들을 더 자세히 설명했으니, 꼼꼼히 살펴보시면 분명 좋은 결과가 있을 것입니다!

I. 단원 지도 계획 및 성취 기준

이 교과서는 컴퓨팅을 활용한 문제 해결 능력 함양을 목표로 합니다. 특히, 데이터를 효율적으로 관리하고 보호하기 위한 데이터 압축과 암호화 기술을 깊이 있게 다룹니다. 또한, 현대 사회에서 중요성이 커지고 있는 빅데이터의 개념과 특징을 이해하고, 이를 수집, 처리, 관리, 시각화하는 기술적인 측면을 강조합니다. 무엇보다 중요한 것은 이러한 데이터 분석 과정에서 발생할 수 있는 사회적, 윤리적 문제를 인식하고, 편향되지 않은 투명한 분석을 통해 올바른 결론을 도출하는 태도를 함양하는 것입니다.

성취 기준 상세**:**

- **[12**정**02-01]** 디지털 데이터 압축의 개념과 필요성 이해 및 효율성 분석**/**평가**:** 디지털 데이터를 압축하는 원리를 이해하고, 손실 압축과 무손실 압축의 차이를 명확히 구분하며, 각 압축 방법의 효율성을 특정 상황에 적용하여 평가할 수 있어야 합니다. (예: 어떤 데이터에 어떤 압축 방식이 유리한지 판단)
- **[12**정**02-02]** 암호화 개념 이해 및 데이터를 안전하게 관리하는 사례 비교**/**분석**:** 암호화의 정의, 평문/암호문, 암호화/복호화 키의 개념을 정확히 이해하고, 대칭키와 비대칭키 암호화 방식의 작동 원리를 비교하며, 일상생활 및 실제 시스템에서 암호화가 어떻게 데이터를 보호하는지 구체적인 사례를 들어 설명할 수 있어야 합니다. (예: HTTPS, 공동 인증서, 블록체인 등)
- **[12**정**02-03]** 빅데이터 개념과 특징 이해를 바탕으로 문제 해결에 적합한 데이터 수집**:** 빅데이터의 5V(Volume, Velocity, Variety, Veracity, Value) 특징을 명확히 설명하고, 주어진 문제 해결을 위해 어떤 유형의 데이터를 어떤 방법(센서, 설문, 공공/민간 데이터)으로 수집해야 하는지 계획하고 실행할 수 있어야 합니다. 데이터 전처리의 중요성도 이해해야 합니다.
- **[12**정**02-04]** 빅데이터 분석 도구를 활용한 데이터 시각화 및 의미**/**가치 해석**:**

다양한 데이터 시각화 도구(막대, 선, 원, 히스토그램, 산점도, 박스 플롯 등)의 특징과 용도를 이해하고, 실제 데이터를 목적에 맞게 시각화할 수 있어야 합니다. 시각화된 데이터가 나타내는 의미와 숨겨진 가치를 논리적으로 해석하고 설명할 수 있어야 합니다.

01. 데이터 압축 **(**교과서 **44~51**쪽**)**

1. 디지털 데이터 압축의 개념과 필요성

- 개념**:** 데이터 압축은 디지털 데이터의 크기를 특정 규칙(알고리즘)에 따라 줄이는 기술입니다. 마치 이불을 진공 압축하거나 캔을 찌그러뜨려 부피를 줄이는 것과 같은 원리입니다.
- 필요성**:**
  - 데이터 전송 속도 개선**:** 파일 크기가 작아지므로 네트워크를 통한 업로드 및 다운로드 시간이 단축됩니다.
  - 저장 공간 절약**:** 제한된 저장 장치(하드디스크, 클라우드 등)에 더 많은 데이터를 보관할 수 있습니다.
  - 컴퓨팅 시스템 자원 효율적 활용**:** 적은 자원으로도 데이터 처리 및 관리가 가능해집니다.
  - 인터넷 사용량 제한 시 데이터 전송 비용 절감**:** 데이터 사용량에 따라 요금이 부과되는 환경에서 유리합니다.
  - 디지털 자원 및 전기 에너지 절약**(**탄소 중립 기여**):** 데이터 전송 및 저장에 필요한 에너지 소비를 줄여 환경 보호에 기여합니다. (예: 100MB 파일을 50MB로 압축하면 약 50%의 에너지 절감 효과)

2. 데이터 압축 방법의 종류

- 크게 손실 압축과 무손실 압축으로 나뉘며, 복원 시 원본과의 일치 여부가 가장 큰 차이점입니다.
  - 무손실 압축**:**
    - 특징**:** 원본 데이터의 모든 정보(bit for bit)를 유지하며 압축합니다. 데이터 손실이 전혀 없습니다.
    - 압축률**:** 비교적 낮지만, 데이터 품질은 최상으로 유지됩니다.

        - 복원**:** 압축 해제 시 복원된 데이터는 원본 데이터와 100% 일치합니다.
        - 주요 사용처**:** 텍스트 파일, 프로그램 코드, 압축이 덜 중요한 일부 이미지(로고, 아이콘 등), 원본 품질 유지가 필수적인 오디오/비디오 아카이빙.
        - 예시**:** PNG, GIF (이미지), WAV, FLAC (소리), ZIP, RAR (일반 파일 압축).
    - 손실 압축**:**
        - 특징**:** 사람이 인지하기 어렵거나 중요도가 낮은 일부 정보를 제거하여 압축합니다. (예: 이미지의 미묘한 색상 변화, 소리의 초고음/미세음역대).
        - 압축률**:** 비교적 높고 파일 크기를 크게 줄일 수 있습니다.
        - 복원**:** 압축 시 손실이 발생했기 때문에 복원된 데이터는 원본 데이터와 완전히 일치하지 않습니다.
        - 주요 사용처**:** 대용량 동영상, 사진, 웹 전송용 이미지, 스트리밍 오디오 등 품질 저하가 어느 정도 허용되는 경우.
        - 예시**:** JPEG (이미지), MP3, WMA (소리), MPEG, H.264 (비디오).
- 압축률 공식**:** (원본 데이터 크기 - 압축된 데이터 크기)/원본 데이터 크기×100(%)
    - 예시**:** 원본 100KB 데이터가 30KB로 압축되었다면, 압축률은 (100−30)/100×100=70% 입니다. 압축률이 높을수록 데이터를 더 많이 줄인 것입니다.
- 선택 기준**:** 데이터 압축 방법은 데이터의 유형 (텍스트 파일, 동영상 등)과 사용 목적 (원본 정보 유지 여부)에 따라 적절히 선택해야 합니다.

3. 데이터 유형별 압축 방법

- 문자열 데이터 압축**:** 문자가 연속적으로 나열된 데이터를 이진 코드(비트 또는 바이트)로 표현하며, 일반적으로 한 문자당 8비트(1바이트)를 사용합니다.
    - 런**-**렝스**(Run-length)** 압축**:** 동일한 문자가 연속해서 반복될 때, 해당 문자와 반복 횟수를 함께 기록하는 방식입니다.
        - 효과적인 경우**:** 아이콘, 로고 등 단색 영역이나 연속된 값이 많은 이미지, 혹은 텍스트에서 반복되는 문자열.
        - 예시**:** 'aaabbccccc' (10개의 문자, 80비트) → 'a3b2c5' (6개의 문자, 48비트)로 압축률 40%

- 허프만**(Huffman)** 압축**:** 문자열에서 자주 나타나는 문자에는 더 짧은 코드를, 드물게 나타나는 문자에는 긴 코드를 할당하여 전체 비트 수를 줄입니다.
  - 효과적인 경우**:** 문자 빈도가 불균등한 모든 종류의 텍스트 데이터.
  - 과정**:**
    1. 문자열 내 각 문자의 빈도수를 계산하고 정렬합니다.
    2. 각 문자와 빈도수를 저장하는 '노드'를 생성합니다.
    3. 빈도수가 가장 작은 두 노드를 선택하여 결합(합쳐서 부모 노드 생성)하고, 이 과정을 모든 노드가 하나의 '뿌리 노드(루트 노드)'로 연결될 때까지 반복하여 '이진 트리'를 만듭니다.
    4. 이진 트리의 뿌리 노드에서 각 자식 노드까지의 경로를 따라 이동하며 이진 코드(왼쪽 0, 오른쪽 1)를 생성합니다.
    5. 원본 문자열을 생성된 이진 코드로 변환하여 표현합니다.
  - 예시**:** 'abracadabra' (11개 문자, 88비트)를 허프만 압축 시 23비트로 압축 가능 (압축률 약 73.9%).
- 효율성 비교**:** 런-렝스 압축은 연속적으로 반복되는 문자에 효율적이고, 허프만 압축은 연속적이지 않아도 자주 나타나는 문자에 효율적입니다. 문자열 특성에 따라 적절한 압축 방식을 선택하는 것이 중요합니다.

- 이미지 데이터 압축**:** 이미지는 픽셀(화소)이라는 작은 점들이 모여 구성됩니다. 이미지 데이터 압축은 이미지 내에서 반복되는 패턴을 저장하거나, 색상 정보를 줄이는 등 다양한 방법을 사용합니다.
  - 손실 압축 **(JPEG):**
    - 특징**:** 이미지의 색상 정보를 줄이거나, 사람이 잘 인식하지 못하는 일부 이미지를 제거하여 압축합니다. 압축률이 높아 파일 크기가 매우 작아집니다.
    - 단점**:** 압축률이 높을수록 이미지 품질이 낮아질 수 있으며, 이미지를 확대할 경우 경계 부분이 매끄럽지 않은 '계단 현상(aliasing)'이 발생할 수 있습니다.
    - 적합한 용도**:** 웹페이지 이미지, 이메일 첨부 파일 등 파일 크기가 중요한 경우.
  - 무손실 압축 **(PNG, GIF):**
    - 특징**:** 이미지 품질이 손상되지 않고 원본 그대로 유지됩니다. 파일 크기는 상대적으로 큽니다.

        - **PNG:** 다양한 색상(풀 컬러)과 투명도, 반투명도를 지원하여 고품질의 그래픽, 웹 아이콘, 로고 등에 적합합니다.
        - **GIF:** 256가지 색상(8비트) 표현이 가능하며, 간단한 애니메이션(움직이는 이미지) 작업에 특히 적합합니다.
    - **WebP:** 구글에서 개발한 최신 파일 형식으로, 손실 및 무손실 압축을 모두 지원합니다. 동일 품질에서 JPEG나 GIF보다 더 작은 용량을 차지하며 투명도를 지원하는 장점이 있어 웹 최적화에 유리합니다.
    - 참고**:** 비트맵(픽셀 기반)과 벡터(수식 기반) 이미지의 차이도 이해해야 합니다.
- 소리 데이터 압축**:** 원본 소리 데이터에서 사람의 청각에 상대적으로 영향을 덜 미치는 부분(초고음 대역, 잘 들리지 않는 작은 소리)이나 덜 중요한 정보를 제거하여 파일 크기를 줄입니다.
    - 손실 압축 **(MP3, WMA):**
        - 특징**:** '지각 부호화(Perceptual Coding)' 원리를 이용하여 사람이 잘 인지하지 못하는 소리 성분을 제거합니다. 파일 크기가 매우 작아집니다.
        - 적합한 용도**:** 웹 스트리밍, 휴대용 장치, 음악 다운로드 서비스 등.
        - 비트 전송률**(Bitrate):** 초당 소리를 전달하는 비트 수. 비트 전송률이 높을수록 음질이 좋지만 용량이 증가합니다. (예: MP3는 다양한 비트 전송률 지원)
    - 무손실 압축 **(WAV, FLAC):**
        - 특징**:** 원본 소리 데이터의 품질을 완벽하게 유지합니다. 파일 크기가 커서 저장 공간을 많이 차지합니다.
        - 적합한 용도**:** 고품질 음악 감상, 전문적인 오디오 편집, 아카이빙.
        - **FLAC:** 음원 사이트에서 고음질 음악 제공 시 많이 사용되는 무손실 압축 형식입니다.
    - 품질 비교 요소**:**
        - 샘플링 주파수**(Sampling Frequency):** 초당 표본화 횟수. 높을수록 원음에 가깝지만 데이터 양 증가.
        - 비트 깊이**(Bit Depth):** 표본화된 소리의 높낮이를 나타내는 비트 수. 클수록 음질 개선.

        - 스펙트로그램**(Spectrogram):** 시간에 따른 주파수 영역의 변화를 시각화한 그래프로, 소리 신호의 스펙트럼 변화를 직관적으로 관찰할 수 있습니다.

02. 데이터 암호화 **(**교과서 **52~59**쪽**)**

1. 암호화 개념

- 암호화**:** '평문(Hello)'과 같은 원본 데이터를 '암호화 알고리즘' 및 '암호화 키'를 이용하여 '암호문(011011...)'과 같이 제삼자가 이해하기 어려운 변형된 형태로 바꾸는 보안 기술입니다. 허락되지 않은 사람이 데이터를 보더라도 내용을 알 수 없게 만듭니다.
- 복호화**:** 암호화된 데이터를 다시 원본 데이터로 되돌리는 보안 기술입니다. 이때 '복호화 키'가 사용됩니다.
- 암호화 키**/**복호화 키**:** 암호화 및 복호화 과정에 사용되는 수학적으로 표현된 규칙의 집합입니다.
- 필요성**:** 스포츠 경기의 수신호처럼, 컴퓨팅 시스템에서도 데이터를 안전하게 전달하고 보호하기 위해 암호화 과정이 필수적입니다.

2. 암호화 방식의 종류

- 데이터를 암호화하고 복호화하는 과정에서 사용하는 키의 방식에 따라 크게 두 가지로 나뉩니다.
    - 대칭키 암호화 방식 **(**비밀키 암호화 방식**):**
        - 특징**:** 암호화와 복호화 과정에서 **동일한 키(비밀키)**를 사용합니다.
        - 작동 원리**:** 송신자와 수신자가 미리 비밀키를 교환하여 공유하고 있어야 합니다. 이 키를 모르면 암호문을 해독할 수 없습니다.
        - 장점**:** 암호화/복호화 속도가 빠릅니다.
        - 단점**:** 통신하는 모든 당사자가 비밀키를 공유해야 하므로, 키 관리 및 분배가 복잡하고 안전성이 떨어질 수 있습니다 (키가 탈취되면 모든 데이터가 위험).
        - 적합한 용도**:** 대량의 데이터를 빠르게 암호화/복호화해야 하는 경우 (예: 파일 암호화, 대용량 통신).
    - 비대칭키 암호화 방식 **(**공개키 암호화 방식**):**
        - 특징**:** 암호화와 복호화 과정에서 **서로 다른 두 개의 키 (공개키와

개인키)**를 사용합니다. 이 두 키는 수학적으로 쌍을 이룹니다.

- 작동 원리**:**
  - 공개키 **(Public Key):** 누구나 알 수 있도록 공개되어 있습니다. 송신자는 수신자의 공개키를 이용해 데이터를 암호화합니다.
  - 개인키 **(Private Key):** 특정 개인(수신자)만이 소유하고 비밀리에 보관합니다. 공개키로 암호화된 암호문은 해당 개인키로만 복호화할 수 있습니다.
- 장점**:** 키 분배가 용이하고, 키 노출 위험이 적어 안전성이 높습니다.
- 단점**:** 암호화/복호화 속도가 대칭키 방식보다 느립니다.
- 적합한 용도**:** 키 교환이 어렵거나 높은 보안 신뢰도가 요구되는 경우 (예: 디지털 서명, 인증서, 암호화된 통신 채널 설정).

3. 암호 기법의 이해 및 적용

- 치환형 암호 기법**:** 평문의 문자를 다른 문자로 '치환'하여 암호문을 만드는 방식입니다. 문자의 순서는 변하지 않고 내용만 바뀝니다. 평문의 각 문자가 암호문의 다른 문자와 1대1로 대응됩니다.
  - 시저 암호**:** 알파벳을 일정 수(키 값)만큼 뒤로 밀어 다른 알파벳으로 치환하는 가장 간단한 형태의 치환 암호 (예: 키 값 18일 때 A → S, B → T). 방법이 간단하여 해독이 쉽다는 단점이 있습니다.
- 전치형 암호 기법**:** 평문의 문자를 재배열하거나, 일정한 패턴에 따라 문자의 '위치'를 변경하여 원래 메시지의 의미를 숨기는 방식입니다. 문자의 내용 자체는 바뀌지 않습니다. (예: 메시지를 표에 세로로 쓰고 가로로 읽는 방식)

4. 암호화 활용 사례

- 개인 정보 보호**:** 스마트폰 생체 인증(지문, 얼굴 인식), 개인 파일 암호화 등을 통해 개인 데이터를 안전하게 관리합니다.
- 기업 및 기관 정보 보호**:** 의료 기관, 금융 기관은 고객 개인 정보를 암호화하고, 교육 기관, 군사 기관, 기업은 기밀/민감 데이터를 암호화하여 보호합니다.
- 안전한 인터넷 통신**:**
  - **HTTPS:** 기존의 HTTP(하이퍼텍스트 전송 규약)에 보안 기능이 확장된

프로토콜입니다. 웹 페이지를 암호화/복호화하여 전송하므로, 로그인 정보(ID, 비밀번호)와 같은 민감한 정보가 해커에게 노출될 위험을 크게 줄여줍니다.
    - 공동 인증서 **(**구 공인 인증서**) /** 민간 인증서**:** 온라인 환경에서 본인 확인과 전자 서명을 가능하게 하는 인증 방식으로, 비대칭키 암호화 방식을 활용합니다. 금융 거래 등 높은 보안이 요구되는 곳에 사용됩니다.
    - **2**단계 인증**:** 비밀번호 외에 스마트폰 인증 코드 등 두 가지 이상의 요소를 조합하여 본인 인증을 강화하는 방식 (예: Google, Naver 로그인).
- 블록체인**(Blockchain):** 데이터를 작은 '블록' 단위로 나누어 암호화된 형태로 이전 블록과 연결(체인)하고, 이 블록들을 네트워크 내의 수많은 컴퓨터(참여자)에 동시에 복제/저장하는 분산원장기술입니다.
    - 특징**:** 데이터가 암호화되어 분산 저장되므로 위변조가 매우 어렵습니다. 조작 시 즉시 들통나 투명성이 높습니다.
    - 활용**:** 암호화폐(비트코인), 분산 신원 증명(DID) 등 다양한 분야에서 보안성과 투명성을 높이는 데 사용됩니다.

03. 빅데이터와 데이터 수집 **(**교과서 **60~65**쪽**)**

1. 빅데이터의 개념과 특징

- 개념**:** 디지털 환경에서 끊임없이 생산되는 대량의 데이터 집합을 의미합니다. 넓은 의미로는 이렇게 대량의 데이터를 빠르게 수집, 처리, 분석하여 유의미한 가치를 추출하고 결과를 분석하는 기술 전반을 포함합니다.
- 빅데이터의 활용 사례**:**
    - 영상 콘텐츠 추천 서비스**:** 사용자의 시청 패턴을 분석하여 맞춤형 콘텐츠를 추천합니다.
    - 유행성 질병 관리 시스템**:** 감염자 현황, 의료 품목 구매량 등을 분석하여 취약 연령대 파악, 필요한 의료 품목 예측 등에 활용됩니다.
    - 축구 경기 감독 시스템**:** 선수와 공의 움직임, 훈련 정보 등을 분석하여 선수 교체 시점 결정 등에 활용됩니다.
- 빅데이터의 특징 **(5V):** 빅데이터의 핵심적인 5가지 특성을 의미하며, 때로는 타당성(Validity)과 휘발성(Volatility)을 추가하여 7V로 확장되기도 합니다.

    - **Volume (규모):** 데이터의 양이 거대합니다. 지금 이 순간에도 다양한 곳에서 엄청난 양의 데이터가 끊임없이 생성되고 있습니다. (예: 온라인 구매 시 구매자의 정보가 카드사, 쇼핑몰에 누적)
    - **Velocity (속도):** 데이터가 생성되는 속도와 처리되는 속도가 매우 빠릅니다. 실시간으로 데이터를 분석하고 반영해야 하는 경우가 많습니다. (예: 내비게이션의 실시간 교통 상황 반영)
    - **Variety (다양성):** 데이터의 형태가 매우 다양합니다.
        - 정형 데이터**:** 표처럼 속성(열)과 값(행)이 명확하게 구분되어 구조화된 데이터 (예: 스프레드시트, CSV 파일). 의미 파악이 쉽고 분석에 바로 활용 가능.
        - 비정형 데이터**:** 고정된 형식이나 구조가 없는 데이터 (예: 동영상, 소셜 미디어 게시물, 음성 기록). 분석 전에 많은 시간과 비용이 드는 '사전 작업'이 필요합니다.
        - 반정형 데이터**:** 고정된 형식은 아니지만, XML, JSON, HTML처럼 어느 정도 구조화된 데이터.
    - **Veracity (정확성):** 데이터의 신뢰성입니다. 데이터의 양이 많아질수록 엉터리 데이터가 섞여 있을 가능성이 높아지므로, 수집된 데이터가 정확한지 확인하는 것이 매우 중요합니다.
    - **Value (가치):** 데이터를 분석하여 얻을 수 있는 유용한 가치입니다. 단순한 데이터 덩어리가 아니라, 분석을 통해 새로운 지식이나 통찰력을 도출할 수 있는 잠재력을 의미합니다.
- 데이터 속성의 종류**:**
    - 범주형 데이터**:** 특정한 범주로 구분되는 데이터.
        - 명목형**:** 단순 분류, 순서 없음 (예: 성별, 계절).
        - 순서형**:** 분류된 값 사이에 순서 존재 (예: 성적, 학년).
    - 수치형 데이터**:** 숫자로 표현되는 측정 가능한 양적 데이터.
        - 연속형**:** 연속적인 값 (예: 온도, 키, 시간).
        - 이산형**:** 연속적이지 않고 셀 수 있는 값 (예: 인원수, 물건 개수).

2. 문제 해결에 적합한 데이터 수집 방법

- 데이터 수집**:** 주어진 문제를 해결하기 위해 필요한 데이터를 모으는 과정입니다. 문제에 필요한 데이터가 무엇인지 파악하고, 기존에 수집된 데이터가 있는지, 직접 수집해야 하는지 등을 고려해야 합니다.
- 다양한 수집 방법**:**
    - 스마트 기기를 활용한 센서 데이터 수집**:** 스마트폰, 스마트워치, IoT(사물 인터넷) 기기 등 내장된 센서를 통해 온습도, 걸음 수, 심박수 등 물리적 환경이나 신체 활동 데이터를 자동으로 수집합니다. (예: '일일 걸음 수가 수면 시간에 영향을 미칠까?' 문제 해결 시 스마트 기기 활용)
    - 설문 조사를 통한 직접 데이터 수집**:** 특정 목적을 위해 설문지, 인터뷰 등을 활용하여 사용자의 의견이나 정보를 직접 수집하는 방법입니다. 자료를 정형화하기 쉽다는 장점이 있습니다. (예: '우리 반 남녀별 선호하는 영화 장르' 파악)
    - 온라인 공공 데이터 또는 민간 데이터 수집**:** 정부 기관(공공 데이터 포털, AI Hub, 기상 자료 개방 포털 등)이나 민간 기업/플랫폼(Kaggle, 네이버 데이터랩, 구글 트렌드 등)에서 공개하는 데이터를 활용합니다. 메타데이터(데이터에 대한 데이터)를 참고하여 목적에 맞는 데이터를 선택하는 것이 중요합니다.
- 데이터 수집 시 고려 사항 **(**윤리적 측면 포함**):**
    - 수집 가능성**:** 기술적으로 수집이 가능한 데이터인가?
    - 목적 부합성**:** 수집하려는 데이터가 문제 해결 목적에 정확히 부합하는가?
    - 개인 정보 보호 및 사생활 침해 문제**:** 개인을 식별할 수 있는 정보(이름, 전화번호 등)는 최소한으로 수집하고, 꼭 필요한 경우 정보 제공자로부터 동의를 받아야 합니다. 프로젝트가 끝나면 반드시 파기해야 합니다.
    - 데이터 출처와 투명성**:** 데이터의 출처가 정확하고 신뢰할 수 있는지 확인해야 합니다. 분석 과정도 투명하게 공개되어야 합니다.
    - 데이터 소유권과 저작권**:** 데이터를 사용할 때 저작권 정책을 준수해야 합니다.
    - 공정성 및 차별 금지**:** 편향된 데이터를 사용하면 편향된 분석 결과가 도출되어 특정 집단에 대한 차별이 발생할 수 있습니다. 데이터가 편향되지 않고 분석 과정이 투명해야 합니다.
    - 비용 적절성**:** 데이터 수집 및 관리에 필요한 비용이 적절한가?

- 데이터 전처리**(Data Preprocessing):** 수집한 데이터는 종종 오류, 누락, 중복 등이 포함되어 분석에 부적합한 경우가 많습니다. 데이터 전처리는 데이터를 분석 가능한 상태로 정제하고 가공하는 필수적인 과정입니다.
    - 주요 과정**:**
        1. 불필요한 속성 제거**:** 분석에 필요 없는 열(컬럼)을 삭제합니다.
        2. 데이터 형식 오류 제거**:** 같은 값이라도 형식이 다른 경우(예: '1학년'과 '1')를 통일합니다.
        3. 이상값 처리**:** 중복되거나 다른 데이터와 비교하여 과도하게 벗어난 값(Outlier)을 제거하거나 평균값 등으로 대체합니다.
        4. 결측값 처리**:** 손실되거나 누락된 값(Missing Value)을 제거하거나 평균값 등으로 대체합니다.
        5. 데이터 통합**:** 여러 출처의 데이터를 하나로 합칠 때, 공통 속성을 기준으로 형식과 범위를 통일합니다.

04. 데이터 시각화와 해석 **(**교과서 **66~74**쪽**)**

1. 데이터 시각화의 개념

- 개념**:** 데이터의 속성(값)을 그래프, 차트 등 도형, 선, 색과 같은 시각적 요소를 활용하여 사람들이 더 쉽게 이해하고 파악할 수 있도록 시각적으로 표현하는 기술입니다.
- 필요성**:**
    - 데이터 패턴 파악 용이**:** 복잡한 대량의 데이터 속에서 규칙성, 추세, 이상치 등을 한눈에 파악할 수 있습니다.
    - 데이터 간 관계 및 미래 예측 도움**:** 여러 데이터 속성 간의 상관관계나 인과관계를 파악하고, 이를 바탕으로 미래를 예측하는 데 도움을 줍니다.
    - 주장의 핵심 명확화**:** 데이터를 기반으로 한 시각화 자료는 발표나 설득 시 자신의 주장을 다른 사람이 명확하고 직관적으로 이해하도록 돕습니다. (예: 공약 발표 시 원그래프 활용)
    - 비정상 데이터값 식별**:** 입력된 대량의 데이터에서 오류나 이상한 값을 쉽게 찾아낼 수 있습니다. (예: 심박수 0인 데이터)

2. 데이터 시각화 종류

- 시각화의 목적(무엇을 보여주고 싶은가?)과 데이터의 유형(어떤 종류의 데이터인가?)에 따라 가장 적절한 시각화 방법을 선택해야 합니다.
- 명확한 시각화를 위한 고려 사항**:**
  1. 데이터 형상화**:** 데이터의 특성을 가장 잘 설명할 수 있는 방법으로 시각화합니다.
  2. 시각적 요소 고려**:** 시각적 표현을 돕는 색상, 글꼴, 레이블, 범례 등 세밀한 요소들을 효과적으로 사용합니다.
- 주요 시각화 종류 및 특징**:**
  - 막대그래프**(Bar chart):**
    - 특징**:** 막대의 길이 또는 높이를 이용하여 시각화합니다.
    - 용도**:** 개수나 수치를 비교할 때 가장 효과적입니다. 세로 또는 가로 형태로 표현할 수 있습니다. (예: 학년별 인원수, 식품 1kg당 온실가스 배출량 비교)
  - 선 그래프**(Line chart):**
    - 특징**:** 점과 점을 선으로 연결하여 시각화합니다.
    - 용도**:** 시간에 따른 데이터의 변화 흐름이나 추이를 추적하고, 여러 데이터 값의 변화를 비교할 때 효과적입니다. (예: 특정 인물 소셜미디어 언급량 추이, 지하철 승객 수 변화)
  - 박스 플롯**(Box plot):**
    - 특징**:** 데이터로 얻어낸 통계적 수치(최솟값, 25%, 50%(중앙값), 75%, 최댓값)를 박스 모양으로 표현합니다. 위아래로 뻗은 선(수염)은 데이터의 범위를 나타내며, 박스 바깥의 점은 '이상치'를 의미합니다.
    - 용도**:** 데이터 집합의 분포 범위와 통계적 특성을 시각적으로 확인하고, 이상치**(**과도하게 벗어난 값**)**를 발견하는 데 매우 용이합니다. 여러 박스 플롯을 나란히 두어 그룹 간 분포를 비교할 수 있습니다. (예: A시 7월 일평균 기온 분포 비교, 지하철역별 승차 승객 수 분포)
  - 히스토그램**(Histogram):**
    - 특징**:** 가로축에 정량화된 특정 간격(계급)을 표시하고, 각 구간에 해당하는 값의 빈도(개수)를 세로축에 막대 형태로 시각화합니다. 모든 막대가 공백 없이 붙어 있습니다.

        - 용도**:** 데이터 속성값의 전체적인 분포 형태를 파악할 때 효과적입니다. 데이터가 어느 구간에 집중되어 있는지, 어떤 모양의 분포를 이루는지 시각적으로 보여줍니다. (예: 연령별 인구 분포)
        - 막대그래프와의 차이점**:** 막대그래프는 범주 간 크기 비교가 주 목적이지만, 히스토그램은 연속적인 데이터의 분포 형태를 시각화하는 것이 주 목적입니다.
    - 산점도**(Scatterplot):**
        - 특징**:** 두 개(또는 그 이상)의 데이터 속성값을 각각 x축과 y축에 놓고, 해당 값을 가지는 위치에 점으로 표현합니다.
        - 용도**:** **두 데이터 속성 간의 관계나 연관성(상관관계)**을 확인하는 데 효과적입니다. 점들이 특정 방향으로 모여 있으면 상관관계가 있다고 판단할 수 있습니다. (예: 기온에 따른 공공 자전거 대여 건수)
    - 원그래프**(Pie chart):**
        - 특징**:** 원형을 여러 부채꼴 조각으로 나누어 표현합니다. 각 조각의 크기는 전체에 대한 해당 항목의 비율을 나타냅니다.
        - 용도**:** 전체에 대한 각 속성의 비율을 비교할 때 효과적입니다. (예: 매점 간식 판매량 비율)

3. 데이터 시각화 결과 해석

- 데이터를 시각화하는 것만큼이나, 시각화된 데이터가 담고 있는 의미와 가치를 올바르게 해석하고 타인에게 공유하는 과정이 매우 중요합니다.
- 주의사항**:** 동일한 데이터를 가지고도 수집 및 전처리 과정의 편향성, 분석 방법, 심지어 해석하는 관점에 따라 다른 분석 결과와 해석이 나올 수 있습니다. 따라서 분석 과정의 투명성과 논리성을 확보하는 것이 중요합니다.
- 효과적인 해석을 위한 노력**:**
    - 데이터 수집 및 처리 과정에서 편향성을 최소화했는지 확인합니다.
    - 전반적인 분석 과정이 논리적이고 효율적이었는지 되짚어봅니다.
    - 다른 사람의 분석 결과를 공유하고, 다양한 의견을 비교하며 데이터를 폭넓게 이해하려고 노력합니다.
- 이러한 과정을 통해 이전에는 발견하지 못했던 새로운 사실을 발견하고, 데이터에

대한 **통찰력(Insight)**을 기르며, 최종적으로 신뢰할 수 있는 의사 결정을 내리는 데 도움을 받을 수 있습니다.

대단원 마무리 문제 및 해설 **(**교과서 **78~79**쪽**)**

문제를 풀면서 이 단원에서 배운 내용을 정리해 보세요.

01. 데이터 압축에 대한 설명으로 옳지 않은 것은?
① 이미지 데이터 JPEG 파일 형식은 대표적인 무손실 압축 방식이다.
② 데이터 압축이란 특정한 규칙에 따라 데이터 크기를 줄이는 기술이다.
③ 원본 데이터가 100KB 크기인 데이터가 70KB로 압축된 경우, 압축률은 30%이다.
④ 런-렝스 압축 방법을 사용하면 'AAABB'라는 문자열을 'A3B2'로 압축할 수 있다.
⑤ 손실 압축 방법은 무손실 압축보다 비교적 압축률이 높지만, 복원된 데이터는 원본 데이터와 일치하지 않는다.
[해설**]** JPEG 파일은 손실 압축 방법으로, 이미지의 색상 정보를 줄이거나 일부 이미지를 제거하여 압축합니다.

02. 다음 ( )안에 들어갈 알맞은 말은?
데이터 ( )이란 원본 데이터를 변형된 형태로 바꾸는 보안 기술로, 허락하지 않은 사람에게 데이터가 노출되더라도 데이터의 내용을 확인하기 어렵게 하는 기술이다.
[정답] 암호화
[해설] 암호화는 데이터의 보안 유지를 위해 원본 데이터를 변형하는 것입니다.
03. 치환형 암호 기법을 사용한 암호문을 전달받은 다음 상황에서 **~**에 들어갈 것을 바르게 연결한 것은**?**

암호문 "BDW W VXXW"과 키값 9를 전달받았다. 알파벳을 아래 표처럼 9글자씩 이동한다면, 원본 데이터는 ""일 것이다.

	MNOPQRSTU	VWXYZABCD
①	ABC	N	XZZY
②	SUN	M	MOON
③	SUN	N	MOON
④	XYZ	M	ACCB
⑤	XYZ	N	ACCB

[정답] ③
[해설] 암호문의 각 알파벳과 대응하는 알파벳을 찾아 정리하면, ( ) 안은 각각 SUN, NL MOON이 됩니다.
(B에서 9글자 뒤로 이동하면 S, D에서 9글자 뒤로 이동하면 U, W에서 9글자 뒤로 이동하면 N이 됩니다. 따라서 BDW는 SUN이 됩니다. 같은 방식으로 VXXW를 변환하면 MOON이 됩니다.)
04. 데이터 유형을 바르게 연결한 것은?
① 영화 시간표 → 정형 데이터
② 성적표 → 비정형 데이터
③ 여행 영상 → 정형 데이터
④ 급식 식단표 → 비정형 데이터
⑤ 소셜 미디어 게시물 → 정형 데이터
[정답] ①
[해설] 영화 시간표는 형식이 정해진 데이터로 정형 데이터에 속합니다.
(참고: 성적표는 정형, 여행 영상은 비정형, 급식 식단표는 정형, 소셜 미디어 게시물은 비정형 데이터입니다.)
05. 보기를 보고 프로젝트 주제와 관련하여 필요한 데이터 수집 방법을 바르게 연결한 것은**?**

보기

- 가희**:** 우리 학교 내 동아리 가입 수와 학교생활 만족도의 관계를 분석하고 싶어.
- 나희**:** 나는 요즘 환경 문제에 관심이 많아져서 생활폐기물 현황에 대해 알아보고 싶어.
- 다희**:** 내가 요즘 피곤한 이유를 알아내기 위해 내 수면 시간을 분석하고 싶어.

	센서 데이터 수집	직접 데이터 수집	온라인 공공 데이터 수집
①	가희	나희	다희
②	나희	다희	가희
③	다희	가희	나희
④	다희	나희	가희

⑤	가희	다희	나희

[정답] ③
[해설]

- 가희 **(**동아리 가입 수와 학교생활 만족도**):** 설문 조사를 통한 직접 데이터 수집이 적절합니다.
- 나희 **(**생활폐기물 현황**):** 정부 기관 등에서 제공하는 온라인 공공 데이터 수집이 적절합니다.
- 다희 **(**수면 시간 분석**):** 스마트워치 등 스마트 기기를 활용한 센서 데이터 수집이 적절합니다.

06. 다음 상황에서 가장 적절한 데이터 시각화 방법은?
전기차 판매량이 탄소 중립에 얼마나 영향을 미치는지 분석하기 위해 전기차 판매량과 탄소 배출량의 상관관계를 점으로 표현하여 시각화하려고 한다.
① 막대그래프 ② 산점도 ③ 박스 플롯 ④ 선 그래프 ⑤ 히스토그램

[정답] ②
[해설] 산점도는 두 속성(여기서는 전기차 판매량과 탄소 배출량) 간의 상관관계를 점으로 표현하여 확인하는 데 가장 적합한 시각화 방법입니다.
07. 그래프 설명을 찾아 적어 보자**.**

- 점과 점을 선으로 연결하여, 데이터 변화 흐름을 비교할 때 효과적인 그래프
- 데이터로 얻어 낸 통계적 수치를 박스 모양으로 시각화하는 그래프
- 데이터의 속성값을 점으로 표현하고, 속성 사이 관계를 확인할 때 효과적인 그래프
- 원형의 여러 부채꼴 조각으로 시각화하고, 전체에서 각 속성의 비율을 비교할 때 효과적인 그래프
- 막대의 길이나 높이를 이용하여 시각화하고, 속성값의 개수나 수치를 비교할 때 효과적인 그래프
- 가로축에 특정 간격을 표시한 후, 각 구간 대응 값의 빈도를 세로축에 막대로 시각화하는 그래프

① 막대그래프:	② 선 그래프:
③ 히스토그램:	④ 원 그래프:
⑤ 박스 플롯:	⑥ 산점도:

[정답]

- ① 막대그래프**:** 막대의 길이나 높이를 이용하여 시각화하고, 속성값의 개수나 수치를 비교할 때 효과적인 그래프
- ② 선 그래프**:** 점과 점을 선으로 연결하여, 데이터 변화 흐름을 비교할 때 효과적인 그래프
- ③ 히스토그램**:** 가로축에 특정 간격을 표시한 후, 각 구간 대응 값의 빈도를 세로축에 막대로 시각화하는 그래프
- ④ 원 그래프**:** 원형의 여러 부채꼴 조각으로 시각화하고, 전체에서 각 속성의 비율을 비교할 때 효과적인 그래프
- ⑤ 박스 플롯**:** 데이터로 얻어 낸 통계적 수치를 박스 모양으로 시각화하는 그래프
- ⑥ 산점도**:** 데이터의 속성값을 점으로 표현하고, 속성 사이 관계를 확인할 때 효과적인 그래프

08. 다음 히스토그램에 대한 설명으로 옳은 것은**?**

① 데이터 중 가장 낮은 값은 230이다.
② 데이터의 정확한 평균값은 255이다.
③ 두 번째로 높은 빈도를 보인 값은 200개 이상이다.
④ 255에서 260 사이의 값이 가장 높은 빈도를 보인다.
⑤ 히스토그램으로 시간의 흐름에 따라 각 빈도별 도수가 변화하는 모습을 확인할 수 있다.
[정답] ④
[해설] 히스토그램의 가로축은 특정 구간(계급)을, 세로축은 빈도(도수)를 나타냅니다.

- 이 그래프만으로는 정확한 가장 낮은 값이나 평균값을 알 수 없습니다.
- 가장 높은 빈도를 보인 막대는 255에서 260 사이의 구간이며, 그 빈도는 200개 이상입니다. 하지만 '두 번째로 높은 빈도를 보인 값'은 150 이상 200 미만 구간이므로 ②는 틀렸습니다.

- 히스토그램은 데이터 분포 형태를 보여주는 것이지, 시간의 흐름에 따른 변화를 보여주는 그래프가 아닙니다.

09. '우리 학교 학생들이 선정한 우리 동네 최고의 음식점은**?**'이란 주제를 가지고 데이터 분석 프로젝트를 협업하여 수행하는 친구들의 대화를 보고**,** 데이터 수집 시 주의해야 할 점을 서술하시오**.**

- 산이**:** 우리 학교 친구들 대상으로 음식점별 맛 평점을 수집하는 일은 내가 할게. 이번에 친구들 전화번호도 수집해서, 친구들도 많이 사귀어야지.
- 별이**:** 나는 어제 웹사이트에서 프로젝트 주제와 관련된 개인의 맛집 평점 리스트를 찾았어. 글쓴이의 개인 의견을 기재한 사이트인데, 괜찮겠지?
- 바람이**:** 잘했어 별아. 그런데 맛집 평점을 수집할 때, 최근에 새로 생긴 음식점들은 빼고 진행하는 것이 어떨까? 내 생각에는 새로 생긴 지 얼마 안 되어서 프로젝트에 도움이 되지 않을 것 같아.

[예시 답안**]**

- 산이의 경우 **(**개인 정보 보호**):** 개인 정보(전화번호)는 최소한으로 수집해야 하며, 꼭 필요한 경우 정보 제공자로부터 수집 동의를 받아야 합니다. 프로젝트가 끝난 경우 반드시 파기해야 합니다.
- 별이의 경우 **(**데이터 출처 및 신뢰성**):** 출처가 정확하지 않거나 개인의 주관적인 의견이 강하게 반영된 데이터는 신뢰할 수 없습니다. 명확한 출처의 데이터를 사용하고 데이터의 정확성과 신뢰성을 확인해야 합니다.
- 바람이의 경우 **(**데이터의 편향성 및 공정성**):** 특정 기준(최근 생긴 음식점 제외)으로 데이터를 수집하면 편향된 데이터가 되어, 편향된 분석 결과가 도출되고 특정 집단에 차별이 생길 수 있습니다. 데이터가 편향되지 않고 분석 과정이 투명해야 합니다.

10. 다음 질문을 보고 데이터 분석 프로젝트를 계획해 보자**.**

사람들이 흔히 하는 말 중에 "예전보다 여름이 길어졌다."는 말이 있다. 정말 그럴까?

[예시 답안**]**

- ① 문제 정의**:** 최근 20년 내에 여름이 길어졌는지 과거와 비교하여 분석합니다. (여름의 정의를 명확히 해야 함: 예: 일평균 기온 25도 이상인 날들의 연속 기간)

- ② 데이터 수집 계획**:** 기상청 웹사이트(기상자료개방포털 등)에서 최근 20년 동안의 월별 평균 기온 또는 일별 최고 기온 데이터를 수집합니다.
- ③ 데이터 시각화 계획**:** 과거부터 현재까지의 여름 기간(또는 평균 기온) 변화 추세를 확인할 수 있는 선 그래프를 그리고, 필요에 따라 막대그래프나 박스 플롯을 활용하여 연도별 비교를 추가할 수 있습니다.
- ④ 관련 질문 생각해 보기 **(**추가 탐구**):**
  - 봄과 가을은 짧아졌는가**?** (계절 변화 추이)
  - 우리 학교 방학 기간과 비교해 보거나, 연도별 제일 덥거나 추웠던 기간을 비교해 볼 수도 있습니다.