02 데이터의 구조화와 분석

학습 목표
- 실생활의 데이터를 표, 다이어그램 등의 다양한 형태로 구조화할 수 있다.
- 구조화한 데이터의 관계를 파악하고, 데이터에 기반하여 의미를 해석할 수 있다.

1 데이터 구조화의 필요성

옷장의 옷이나 책상 위의 물건을 종류와 쓰임 등에 따라 정리해 두면 필요한 옷이나 물건을 빠르게 찾을 수 있고, 공간도 효율적으로 활용할 수 있으며 관리하는 데 필요한 시간과 노력을 절약할 수 있다. 컴퓨터의 데이터 파일도 시간 순서나 내용별로 폴더에 정리하면 효율적으로 관리할 수 있다.

이와 같이 **데이터의 종류와 특성, 사용 목적 등에 따라 데이터를 체계적으로 구분하여 정리하는 것을 데이터 구조화라고 한다.** 데이터 구조화는 빠른 데이터 검색과 효율적인 데이터 관리를 가능하게 한다. 또 구조화된 데이터는 시각적으로 표현하기 쉽기 때문에 데이터의 관계 파악과 분석에 활용도가 높다.

[그림 II-13] 다양한 데이터 구조화

2 데이터 구조화의 방법

데이터를 구조화할 때에는 데이터의 의미를 정확히 파악하여 사용 목적에 적합한 형태로 표현해야 한다. 데이터를 구조화하는 방법에는 대표적으로 표와 다이어그램이 있다.

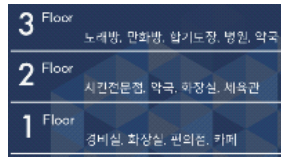
❶ 표

표는 데이터를 기준에 따라 행(가로)과 열(세로)의 2차원 형태로 만든 구조이다. **표는 데이터를 속성별로 구분하여 정리할 수 있고, 데이터의 추가, 삭제, 수정이 편리하여 많은 양의 데이터를 관리할 때 주로 사용한다.** 스프레드시트를 이용하여 쉽게 표로 만들고 데이터를 관리할 수 있다.

표의 구성
표에서 데이터 속성은 열로 표현되고, 가로로 나열된 데이터 값들의 모음인 데이터 단위(레코드)는 행으로 표현된다.

목록
데이터를 한 가지 기준에 따라 일정한 순서로 나열한 구조로 통화 목록, 건물의 층별 안내도 등이 있다.

예 건물의 층별 안내도



수업 시간표

	월	화	수	목	금
1교시	과학	음악	수학	영어	사회
2교시	역사	체육	도덕	한문	역사
3교시	도덕	정보	영어	수학	수학
4교시	정보	국어	역사	과학	과학
5교시	미술	진로	국어	국어	국어
6교시	체육	수학	과학	체육	영어

열

1	A	B	C	D
2	주소록			
3	이름	전화번호	휴대전화 번호	주소
4	김○경	3590-1***	010-3***-****	○○시
5	김○수	3490-2***	010-5***-****	△△군
6	최○주	3695-4***	010-2***-****	☆☆시
7	한○진	6953-8***	010-9***-****	□□시

행

[그림 Ⅱ-14] 표를 이용한 데이터 구조화의 예

해 보기 실생활 데이터를 표로 구조화하기

조사 ‹ 탐구 › 실습 토의 발표
개별

다음은 학생들이 희망하는 동아리 신청 내용을 기록해 놓은 것이다. 표를 이용해 구조화해 보고, 구조화하기 이전과 비교해 어떤 특징이 있는지 써 보자.

이름	1차 신청 동아리	2차 신청 동아리

• 표로 구조화하였을 때의 특징: _____

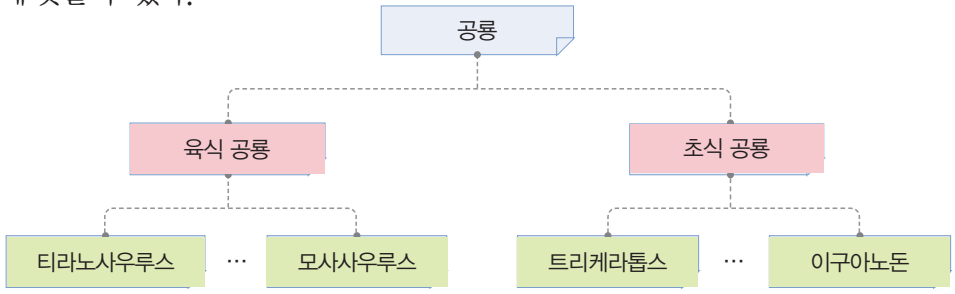
❷ 다이어그램

데이터의 관계
데이터 간의 연결성이나 상호 작용을 말한다.

다이어그램은 점, 선, 도형, 기호 등을 사용하여 데이터의 관계를 표현한 구조이다. 다이어그램은 데이터 간의 관계를 시각적으로 표현하여 이해하기 쉽게 하고 데이터를 단순한 형태로 조직화하여 원하는 데이터를 빠르게 찾을 수 있다. 다이어그램의 종류에는 계층형(tree)과 그래프형(graph)이 대표적이다.

계층형 다이어그램의 예
학급의 조직도, 컴퓨터의 폴더 구조, 책의 목차, 가계도 등이 있다.

계층형 다이어그램 | 계층형 다이어그램은 시작 지점을 중심으로 여러 지점으로 갈라지는 구조로 데이터 간의 분류 체계와 위계 관계를 쉽게 파악할 수 있다는 장점이 있다. 예를 들어, 다음과 같이 공룡의 식성을 기준으로 계층형으로 표현하면 육식 공룡은 공룡의 하위 분류이면서 티라노사우루스, 모사사우루스의 상위 분류로 데이터 간의 상하 관계를 쉽게 이해할 수 있고, 원하는 데이터를 빠르게 찾을 수 있다.

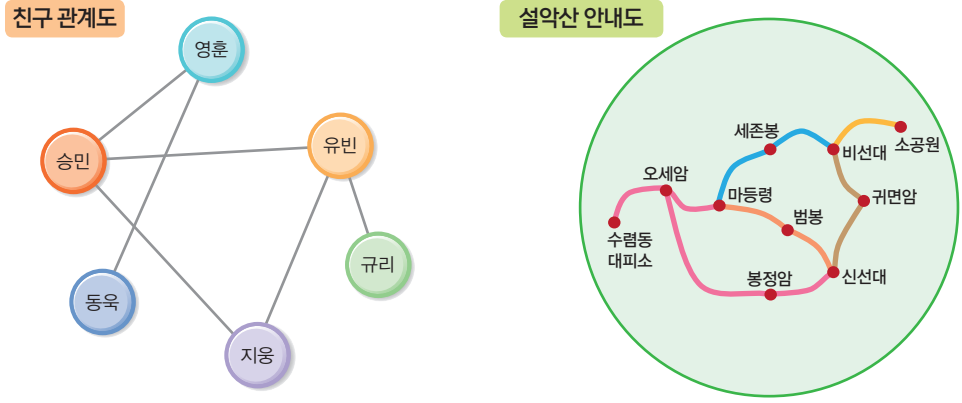


[그림 Ⅱ-15] **계층형 다이어그램**

그래프형 다이어그램의 예
버스와 지하철 노선도, 별자리, 전기 회로도 등이 있다.

그래프형 다이어그램 | 그래프형 다이어그램은 데이터를 나타내는 원 또는 사각형 등의 도형(노드)과 데이터 간의 관계를 나타내는 선(간선)을 이용하여 시각적으로 표현한 것이다. 데이터를 그래프형 다이어그램으로 구조화하면 데이터 간의 복잡한 관계를 직관적으로 파악할 수 있다. 예를 들어 그래프형 다이어그램으로 사회 관계망 서비스(SNS)의 친구 관계를 표현하면 누가 어떤 친구들과 연결되어 있는지 알 수 있고, 복잡한 등산로를 단순하게 표현하여 원하는 등산 경로를 한눈에 파악할 수 있다.

데이터 구조화와 시각화
- 데이터 구조화: 데이터를 체계적으로 정리하고 저장하는 과정으로 데이터의 검색과 관리를 편리하게 한다.
- 데이터 시각화: 구조화된 데이터를 시각적으로 표현하는 것으로 데이터 간의 관계와 패턴을 이해하기 쉽다.



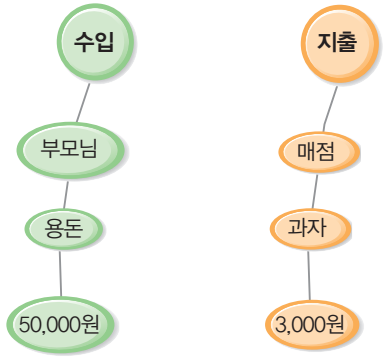
[그림 Ⅱ-16] **그래프형 다이어그램**

해 보기 실생활의 데이터를 구조화하기

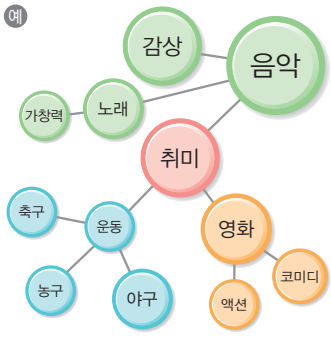
1 다음 상황을 표와 계층형 다이어그램으로 각각 나타내고 특징을 서로 비교해 보자.

예성이는 부모님께 용돈으로 50,000원을 받고, 설거지를 돕고 2,000원을 받았다. 그리고 할머니께 생일 선물로 30,000원을 받고, 할머니 안마를 해드리고 3,000원을 받았다. 이후 예성이는 매점에서 과자 3,000원, 음료 2,000원을 쓰고, 문구점에서 공책 2권 2,000원, 볼펜 3자루 3,000원을 썼다. 또한 영화관에 가서 영화표 8,000원, 팝콘 7,000원을 사용하였다.

수입	부모님	용돈	50,000원
		설거지 도움	2,000원
지출	매점	과자	3,000원
		음료	2,000원



2 그림을 참고하여 '나의 취미와 관심사'를 주제로 서로 관련 있는 단어를 연결하고, 중요도에 따라 글자의 크기를 다르게 하여 그래프형 다이어그램으로 표현해 보자.



3 다음 주제 중 하나를 선택하여 알맞은 방법으로 구조화하고, 그 방법으로 구조화한 이유를 설명해 보자.

주제

국내 여행지
동물의 분류
수학 공식
학생들의 취미와 관심사

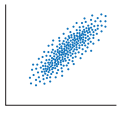
3 데이터 분석

데이터 분석이란 합리적 의사 결정을 위해 수집한 데이터의 의미를 해석하며 유용한 정보를 찾아내는 과정으로, 데이터 수집 및 구조화, 데이터 살펴보기, 그래프로 나타내기, 데이터 의미 해석 등 데이터를 이해하기 위한 일련의 과정을 포함한다.

예를 들어 학습 시간과 성적 사이의 ⊕상관관계를 분석하기 위해 먼저 학생들의 학습 시간과 성적 데이터를 각각 수집하여 표로 구조화할 수 있다. 이어서 데이터 간의 관계를 파악하기 위해 학습 시간과 성적의 관계를 ⊕산점도 그래프로 나타내면 두 속성 사이에 매우 밀접한 상관관계가 있음을 알 수 있다.

⊕ **상관관계**
데이터 속성 간에 서로 어떤 관련이 있는지를 의미한다.

⊕ **산점도 그래프**
데이터를 좌표 평면에 점으로 표시하여 두 데이터 속성 간의 관계를 한눈에 볼 수 있게 해주는 그래프이다. 각 점은 한 쌍의 데이터 값을 나타내며 점의 분포 모양에 따라 다음과 같이 해석할 수 있다.

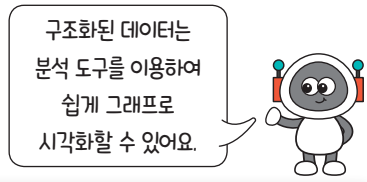


양의 상관관계

- 한 데이터가 증가할 때 다른 데이터도 함께 증가하는 관계이다.
- 점의 분포 모양이 직선에 가까울수록 상관관계가 높다.

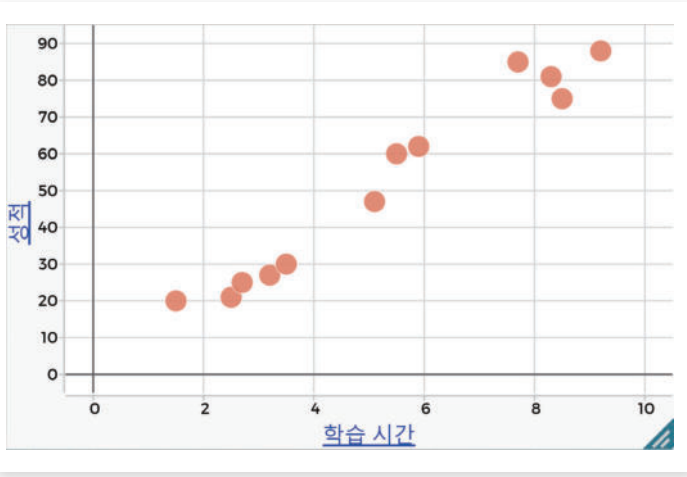
상관관계 없음.

- 두 데이터 간에 일정한 규칙이나 어떤 관계가 없는 경우이다.
- 점의 분포 모양이 무작위로 흩어져 있고, 아무런 규칙이 없다.



학습 시간과 성적
케이스 (12 케이스)

인덱스	학습 시간	성적
1	2.5	21
2	5.1	47
3	3.2	27
4	8.5	75
5	3.5	30
6	1.5	20
7	9.2	88
8	5.5	60
9	8.3	81
10	2.7	25
11	7.7	85
12	5.9	62



[그림 II-17] **학습 시간과 성적의 상관관계 데이터 분석**

4 데이터 간의 관계 파악과 의미 해석

공공데이터포털, ⊕국가통계포털 등의 데이터 전문 웹 사이트를 이용하면 표로 구조화된 데이터를 편리하게 이용할 수 있다. 표를 그래프(도표)와 같이 시각화된 형태로 나타내면 데이터의 관계를 쉽게 이해할 수 있다.

⊕ **국가통계포털 (https://kosis.kr)**
국내 또는 국제 경제, 사회, 문화 관련 통계 자료를 얻을 수 있다.

데이터 수집 및 구조화 | 국가통계포털 웹 사이트의 검색창에 '배출원별 온실가스 배출량: 이산화탄소'를 입력하여 데이터를 검색한다. 조회설정에서 국가와 연도를 각각 다음과 같이 설정한 후 행렬전환하여 연도와 국가별로 가공한 데이터를 ⊕CSV 파일 형식으로 다운로드한다.

⊕ **CSV 파일 형식**
Comma Separated Value의 약자로 데이터 속성을 쉼표로 구분한 문자 데이터 형식이다.

(단위: 이산화탄소 환산톤(천톤))

시점	총 배출량(LULUCF 제외)			
	오스트레일리아	일본	대한민국	미국
2016	410,253.593	1,202,454.571	637,427.680	5,252,932.175
2017	413,655.465	1,186,802.263	650,220.240	5,212,162.345
2018	415,350.721	1,141,668.881	664,976.140	5,377,797.353
2019	415,811.270	1,104,539.821	643,766.990	5,262,145.074

코답을 실행하고, '새 문서'를 클릭한 후 다운로드한 '국가별 이산화탄소 배출량.csv' 데이터 파일을 코답에 끌어 놓으면 다음과 같이 표로 구조화된 데이터가 나타난다.

국가별 이산화탄소 배출량

케이스 (5 케이스)

인덱스	시점	총 배출량 (LULUCF 제외)	총 배출량 (LULUCF 제외)1	총 배출량 (LULUCF 제외)2	총 배출량 (LULUCF 제외)3
1	시점	오스트레일리아	일본	대한민국	미국
2	2016	410253.59	1202454.57	637427.68	5252932.18
3	2017	413655.47	1186802.26	650220.24	5212162.35
4	2018	415350.72	1141668.88	664976.14	5377797.35
5	2019	415811.27	1104539.82	643766.99	5262145.07

데이터 살펴보기 | 연도별 각 국가의 이산화탄소 배출량의 변화를 알아보기 위해 국가별 이산화탄소 배출량을 연도별 데이터로 정리하여 살펴보자.

❶ 클릭하여 '시점'을 '연도'로 바꾼다.

❷ 각각의 셀을 클릭하여 1행의 국가 이름으로 바꾼다.

❸ 클릭하여 '케이스 삭제'를 선택한다.

국가별 이산화탄소 배출량

❶ ❷ *케이스 (5 케이스)*

인덱스	시점	총 배출량 (LULUCF 제외)	총 배출량 (LULUCF 제외)1	총 배출량 (LULUCF 제외)2	총 배출량 (LULUCF 제외)3
❸ 1	시점	오스트레일리아	일본	대한민국	미국
2	2016	410253.59	1202454.57	637427.68	5252932.18
3	2017	413655.47	1186802.26	650220.24	5212162.35
4	2018	415350.72	1141668.88	664976.14	5377797.35
5	2019	415811.27	1104539.82	643766.99	5262145.07

국가별 이산화탄소 배출량

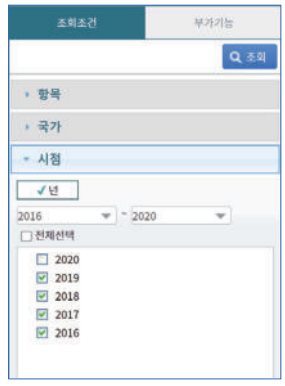
케이스 (4 케이스)

인덱스	연도	오스트레일리아	일본	대한민국	미국
1	2016	410253.59	1202454.57	637427.68	5252932.18
2	2017	413655.47	1186802.26	650220.24	5212162.35
3	2018	415350.72	1141668.88	664976.14	5377797.35
4	2019	415811.27	1104539.82	643766.99	5262145.07

데이터의 수정, 삭제 등의 기능을 이용하면 필요한 속성을 중심으로 정리할 수 있다.

[그림 Ⅱ-18] 정리가 완료된 연도별 각 국가의 이산화탄소 배출량 데이터

[조회 설정]에서 국가와 연도 설정



2020년은 이산화탄소 배출량 데이터가 없는 국가가 있으므로 제외한다.

행렬 전환

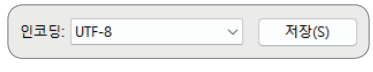
행렬전환 에서 '국가'와 '시점' 위치를 서로 바꿔 준다.

데이터 살펴보기

데이터 관리 기능을 이용하면 원하는 내용을 빠르고 쉽게 찾아볼 수 있다.

코답에서 한글이 정확히 보이지 않을 때 해결 방법

코답에서 CSV 파일을 불러왔을 때 한글이 보이지 않는다면 '메모장'으로 파일을 열고, '다른 이름으로 저장'에서 '인코딩'을 'UTF-8'로 선택하여 저장한다.



코답 메뉴

우측 메뉴 [측정] – [연결 선] 표시하기

꺾은선그래프

자료의 변수가 시간의 경과에 따라 변하는 모습을 나타낼 때 사용한다.

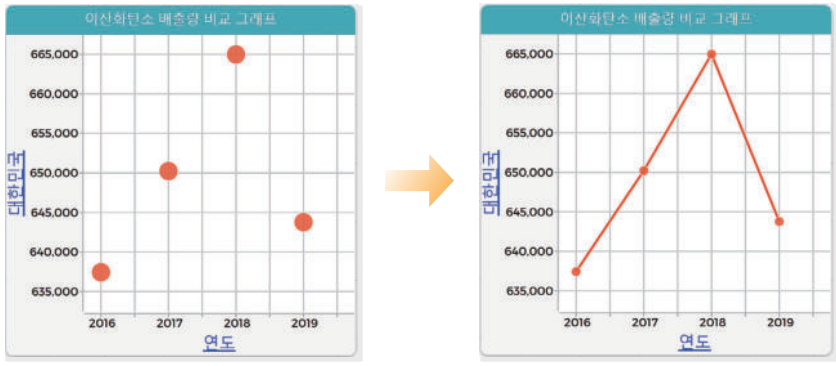
활동 도우미

미국의 이산화탄소 배출량이 다른 세 나라보다 월등히 많아 비교가 어렵지만, 다음과 같이 미국을 제외하면 대한민국·일본·오스트레일리아의 연도별 이산화탄소 배출량 변화를 명확하게 비교할 수 있다.

Tip

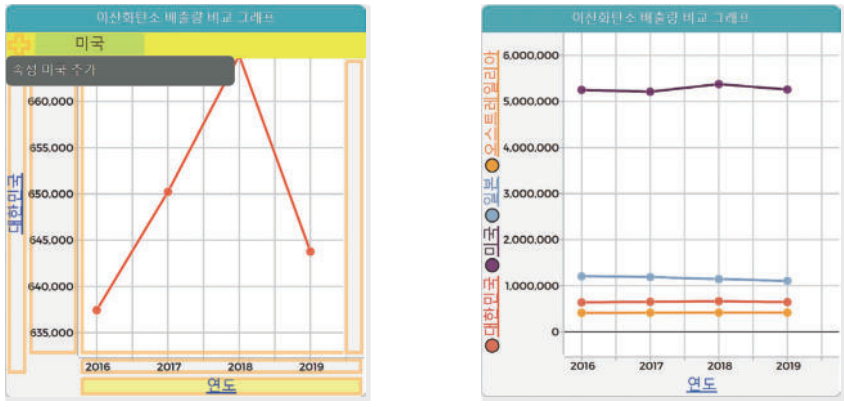
코답에서 나라별 그래프의 선 색깔은 지정할 수 없으므로 각 나라별 그래프 선 색깔은 작성된 그래프마다 달라질 수 있다.

그래프로 표현하기 | 메뉴에서 '그래프' 버튼(그래프)을 클릭하면 표를 그래프로 표현하기 위한 작업 창이 나타난다. 그래프의 [가로축] – [연도]를 선택하고, [세로축] – [대한민국]을 선택한다. 그러면 다음과 같은 형태의 그래프가 나타난다. 우측 메뉴의 측정( )을 클릭하고 '연결 선'을 표시하여 데이터가 표시된 점을 연결하면 다음과 같이 데이터의 변화를 파악할 수 있는 꺾은선그래프가 나타난다.



[그림 Ⅱ-19] **대한민국의 이산화탄소 배출량 비교 그래프**

다른 국가들도 추가하여 연도별 각 국가의 이산화탄소 배출량과 변화를 비교하여 보자. 표에서 ['미국' 속성 이름 드래그] – [그래프 세로축] – [미국 속성 추가]로 미국 데이터를 추가할 수 있다. 일본, 오스트레일리아도 각각 추가하면 다음과 같이 국가별로 이산화탄소 배출량을 비교할 수 있도록 데이터의 변화가 각각 표시된다.



[그림 Ⅱ-20] **국가별 이산화탄소 배출량 비교 그래프**

데이터 의미 해석 | 각 국가의 연도별 이산화탄소 배출량을 2016년부터 2018년까지와 2018년부터 2019년까지로 나누어 비교하면 대체로 이산화탄소 배출량이 줄어든 것을 알 수 있다. 이는 여러 국가가 이산화탄소 배출량을 조절하여 점차 줄이고 있다고 해석할 수 있다. 이처럼 데이터를 수집하고 데이터 간의 관계를 파악하여 의미를 해석하는 데이터 분석 과정을 통해 다양하고 유용한 정보를 얻을 수 있다.

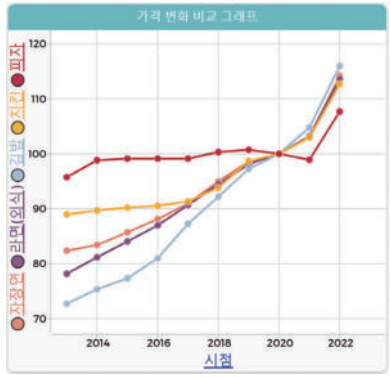
해 보기 경제 분야 데이터 수집·분석하기

다음의 표와 그래프를 참고하여 내가 좋아하는 음식의 물가가 지난 10년 동안 어떻게 변화하였는지 알아보자. 또, 다른 음식과 비교하여 연도별 변화 추이가 어떠한지 분석해 보자.

좋아하는 음식의 가격 변화

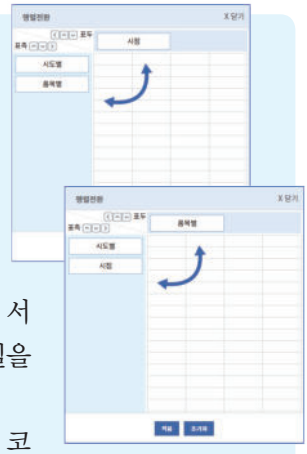
데이터 (10 케이스)

인덱스	시도별	시점	자장면	라면(외식)	김밥	치킨	피자
1	전국	2013	82.35	78.17	72.71	88.97	95.73
2	전국	2014	83.4	81.17	75.36	89.69	98.84
3	전국	2015	85.7	84.03	77.32	90.22	99.12
4	전국	2016	88.11	86.98	80.98	90.53	99.12
5	전국	2017	90.93	90.67	87.26	91.33	99.12
6	전국	2018	94.98	94.46	92.19	93.83	100.32
7	전국	2019	98.59	98.13	97.25	98.67	100.73
8	전국	2020	100	100	100	100	100
9	전국	2021	103.07	103.19	104.78	102.96	98.91
10	전국	2022	114.2	113.56	115.97	112.68	107.69



1 수집한 데이터를 가져오고, 문제 해결 목적에 따라 데이터를 정리하여 데이터 간의 관계를 파악해 보자.

(1) 데이터 수집 및 구조화



❶ 국가통계포털 웹 사이트(https://kosis.kr)에서 검색창에 '품목별 소비자물가지수'를 입력하여 검색하고, 검색 결과에서 '품목별 소비자물가지수(품목성질별: 2020=100)'를 클릭한다.
❷ 조회설정 – [품목별] – [전체 해제]한 후 내가 좋아하는 음식을 검색하고, 시점을 최근 10년(예 2013~2022년)으로 설정한 후 조회를 클릭한다.
❸ [행렬 전환]을 클릭하여 '품목별'과 '시점'을 드래그하여 위치를 서로 바꾼다. 파일 형태를 *.csv, 인코딩 UTF-8로 선택한 후 파일을 다운로드한다.
❹ 코딩을 실행한 후 '새 문서'를 클릭하고 내려받은 데이터 파일을 코딩에 끌어 놓는다.

(2) 그래프로 나타내기

❶ 상단 메뉴의 그래프 – [가로축] – [시점]을 선택한다.
❷ [세로축]을 클릭하고 음식 한 가지를 선택한다. 왼쪽 표에서 나머지 음식도 오른쪽 그래프의 세로축 상단으로 하나씩 끌어 놓는다.
❸ 연도와 물가 지수 숫자 부분을 드래그하여 숫자와 눈금을 적절히 조절한다.

2 그래프를 이용하여 분석한 데이터의 의미를 친구들과 함께 이야기해 보자.

예 최근 10년(2013~2022년) 사이 가격이 가장 많이 오른 음식과 가장 적게 오른 음식은 무엇인지 확인하고, 그 이유에 대해 조사하고 토론해 보자.