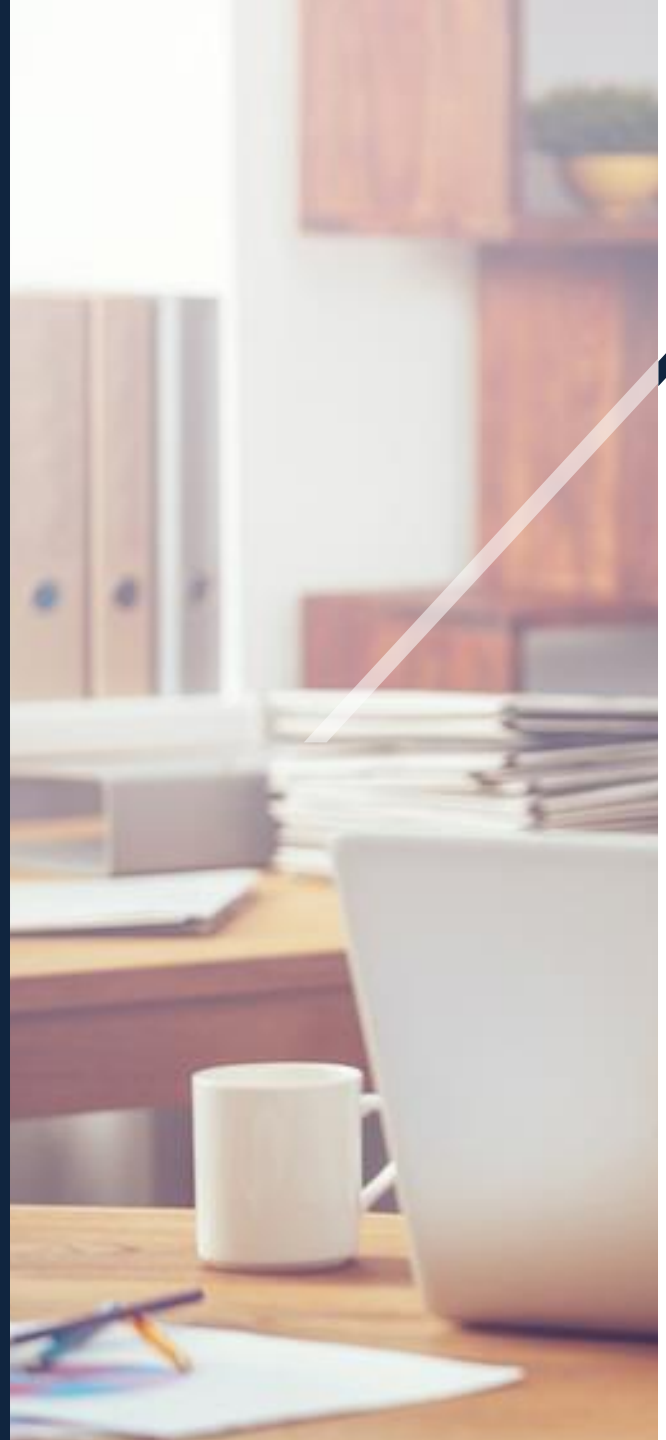
농협경제지주

2023 하반기 엑셀 데이터 분석 교육

(실무 과정 Ⅱ)

Actionable Content
MASO CAMPUS



Actionable Content
MASO CAMPUS

NOTICE

- 본 자료는 저작권법에 의거해 허가 받지 않은 복사, 전재, 편집, 재배포 등을 금지함을 알려 드립니다.

- 본 자료는 온/오프라인 또는 동영상 강의를 전제로 제작되었습니다.
  강사의 부가 설명 없이 본 자료의 내용을 **임의 해석할 경우 잘못된 결론**에 이를 수 있음을 유념하십시오.

- 강의를 캡쳐, 녹음, 녹화하는 등의 콘텐츠의 원천 제공 방식 이외의 **저장 행위를 엄격히 금지**하오니
  필요하신 내용은 수업 틈틈이 개별적으로 메모하시기 바랍니다.

▶ 데이터를 잘 다루기 위한 필수 역량

데이터 기획
데이터 수집, 처리, 분석 과정을
계획, 실행, 개선하는 능력

데이터 수집
필요한 데이터를 선별적으로
추출하고 연결하는 능력

데이터 처리
원본 자료를 분석에 용이한
형태로 변환하는 능력

데이터 분석
다양한 수준의 데이터를 정량적,
정성적으로 분석하는 능력

데이터 시각화
직관적이고 효과적인 방식으로
데이터를 표현하는 능력

▶ Profiling

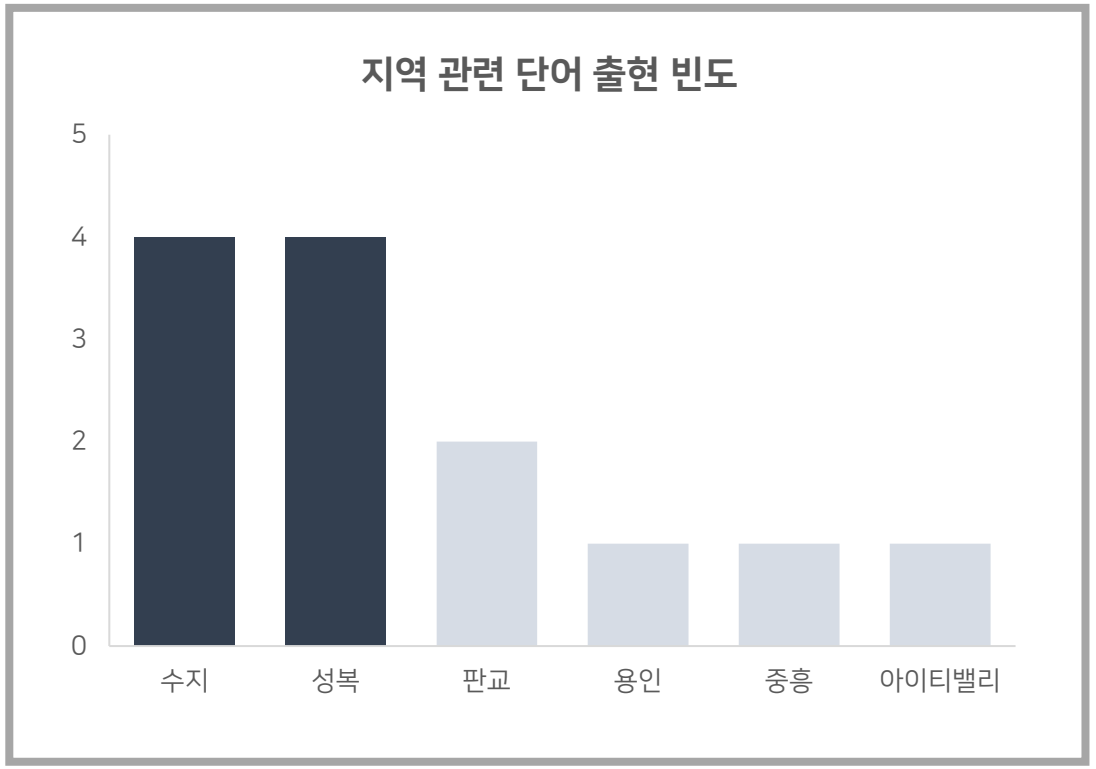
- A씨의 거주 지역(ex. 서울 은평구, 청주시 복대동 등)은 어디입니까?
- A씨의 거주 형태(ex. 아파트, 단독주택, 빌라 등)는 무엇입니까?
- A씨의 동거인은 몇 명(ex. 1~2인 가구, 4인 이상 대가족 등)일까요?
- A씨가 거주하는 아파트는 몇 평(ex. 20평대, 30평대)일까요?
- A씨는 자동차가 있을까요? 있다면 어떤 차일까요?
- A씨의 직장은 어디에 있을까요?

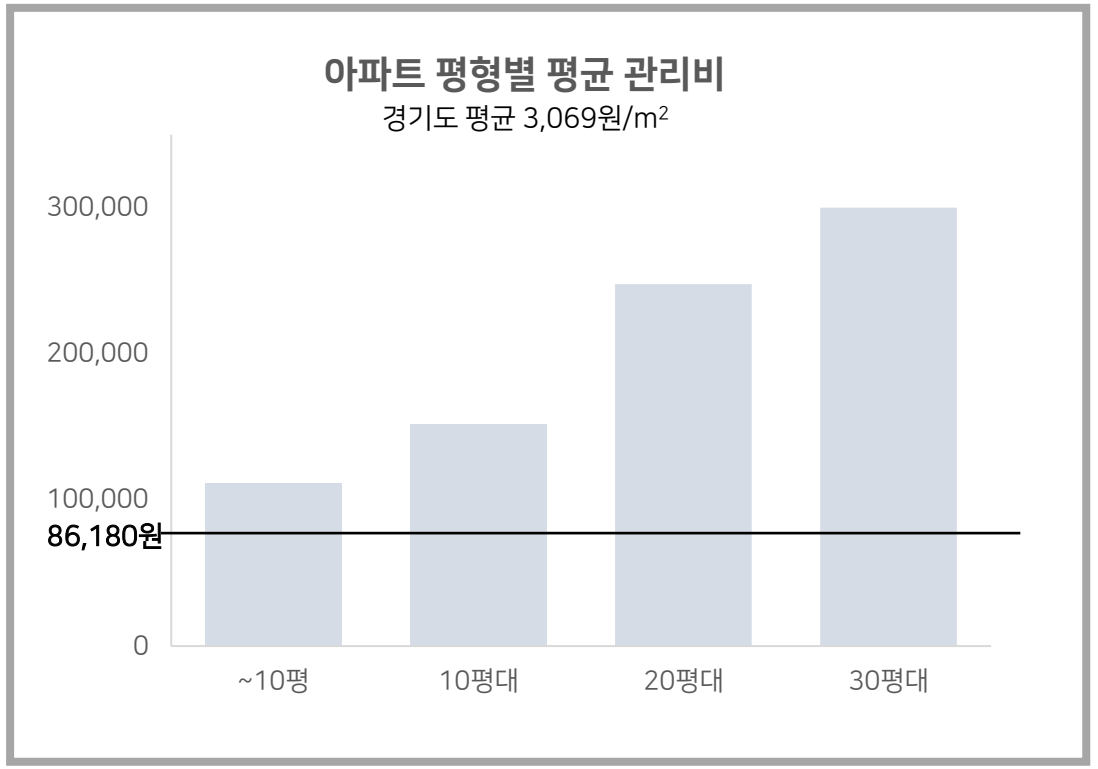
결제승인일자	가맹점명	승인금액(원)
2022-02-01	네이버파이낸셜주식회사	89,100
2022-02-01	솥밭솥밥	47,500
2022-02-01	효성에프엠에스	1,100
2022-02-02	주식회사 디더블유에스	18,000
2022-02-02	주식회사 디더블유에스	7,000
2022-02-03	성복상쾌한이비인후과	4,300
2022-02-04	엘지유플러스	56,750
2022-02-04	삼천리도시가스	2,120
2022-02-05	슈퍼마켓자연숲 중흥점	9,000
2022-02-05	일승상회	25,000
2022-02-06	김밥천국 성복점	6,500
2022-02-07	네이버파이낸셜주식회사	1,352
2022-02-07	카페드롭탑	8,100
2022-02-08	CJ올리브영_온라인몰	24,000
2022-02-08	㈜스타벅스커피코리아	2,960
2022-02-09	플로리다	12,500
2022-02-10	다원수산	20,000
2022-02-11	카카오T대리	26,000
2022-02-11	배스킨라빈스31	15,900
2022-02-12	㈜이마트수지점	113,200
2022-02-13	고봉민김밥	4,000
2022-02-14	Netflix_INICIS 넷플릭스서비스	17,000
2022-02-14	DB손해보험주식회사	32,030
2022-02-14	커피에빠지다	2,000
2022-02-15	다혜수산	15,000

결제승인일자	가맹점명	승인금액(원)
2022-02-15	01월접수 후불하이패스	3,200
2022-02-15	01월접수 후불교통(버스+지하철+통행료)	17,750
2022-02-16	진석수산	15,000
2022-02-17	GS25아이티밸리점	2,950
2022-02-18	11번가	47,000
2022-02-18	공차수지롯데몰점	4,000
2022-02-19	아이디헤어 수지성복점	180,000
2022-02-19	㈜우아한형제들	23,500
2022-02-20	파리바게트	17,100
2022-02-21	㈜스타벅스커피코리아	3,800
2022-02-21	바른약국	26,600
2022-02-22	현대백화점판교점 보테가베네타	975,000
2022-02-22	현대백화점판교점 베즐리	5,900
2022-02-23	세틀뱅크주식회사	24,390
2022-02-23	오토김밥	5,500
2022-02-24	카페드롭탑	8,100
2022-02-24	롯데제과㈜	11,600
2022-02-25	아파트관리비납부 전용 가맹점	86,180
2022-02-25	두나미스(dunamis)	47,000
2022-02-25	애슐리퀸즈수지점	105,700
2022-02-26	주식회사지마켓	22,900
2022-02-26	포인트사용안내 주식회사지마켓	-13,527
2022-02-27	다이소용인성복점	19,100
2022-02-27	CU(씨유)수지	7,350
2022-02-28	메가엠지씨커피	3,500

▶ 어떻게 알았을까?

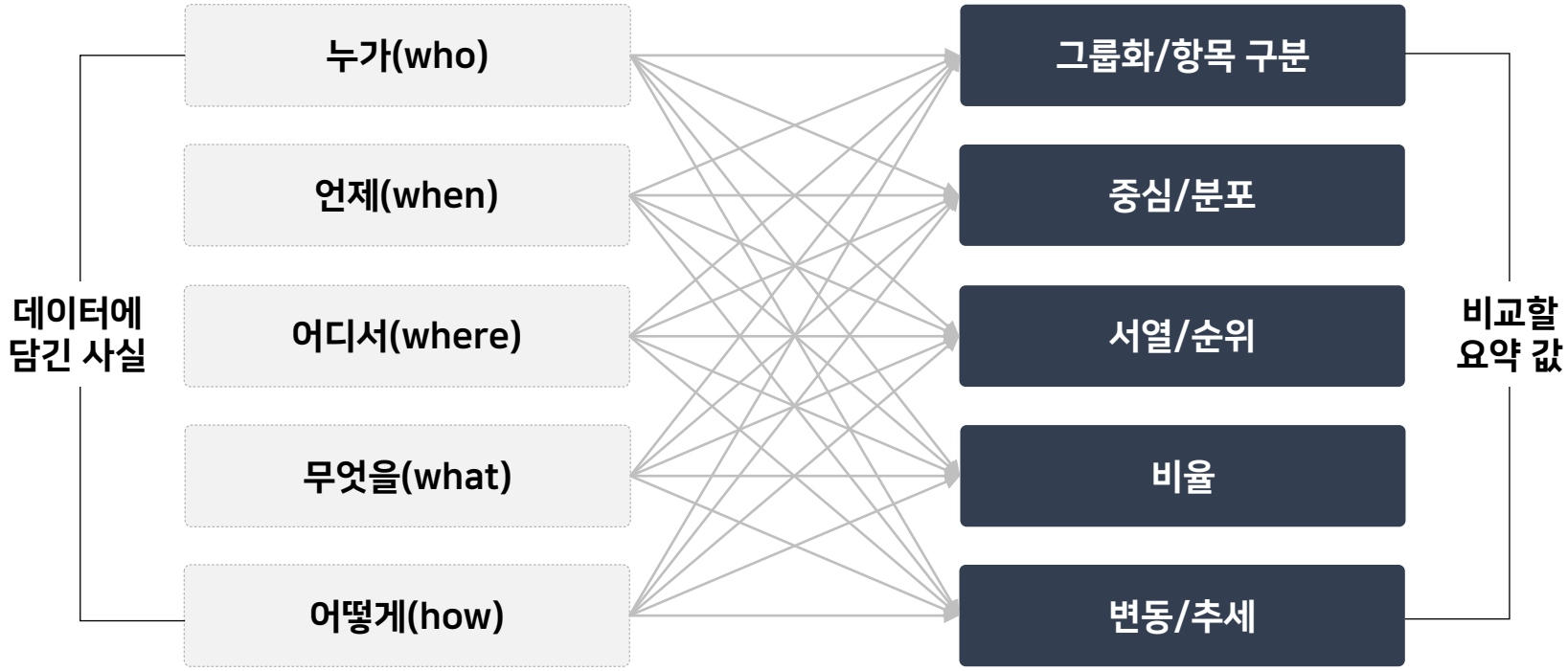
- 데이터 분석이란 데이터를 여러 가지 **값으로 요약** 한 후
  의미있는 여러 집합으로 **모으거나 쪼개어** 집합간에 비교 대조함으로써 이로운 **패턴을 발견**하는 작업이다.



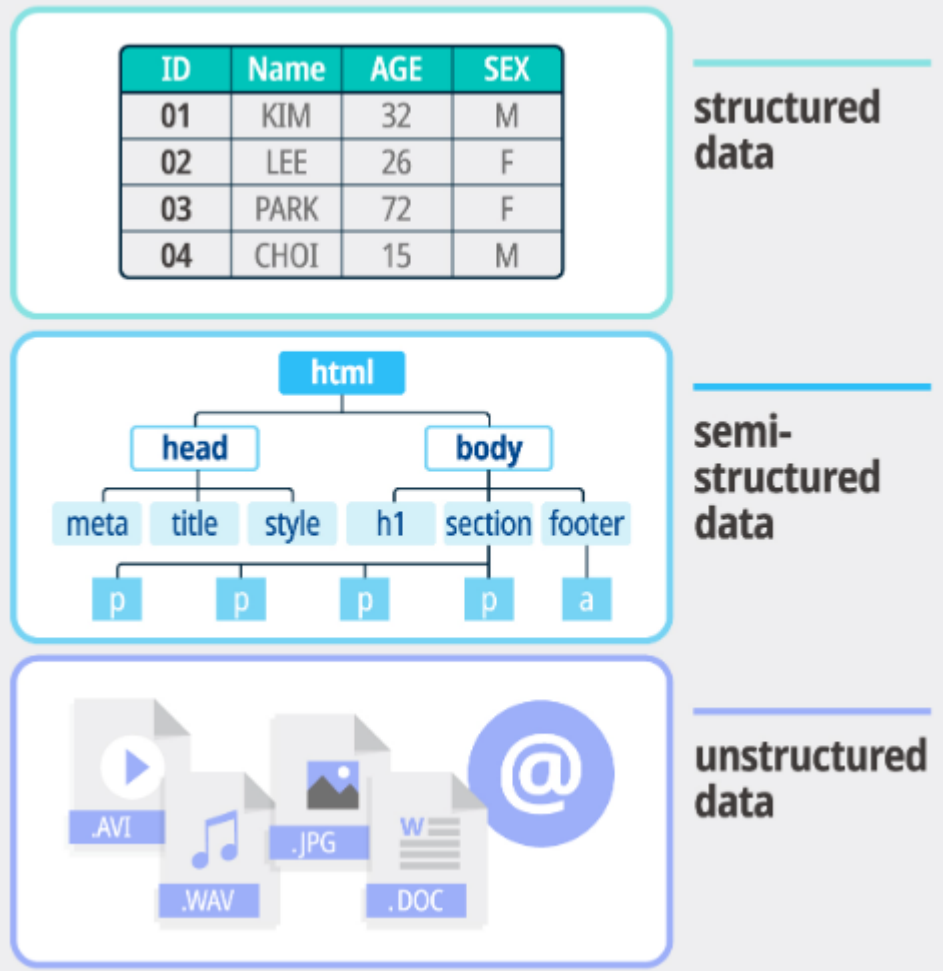


▶ 무엇을 비교하는가?

- 데이터는 현실을 설명하기 위한 자료이므로, 탐색하고자 하는 현실에 대한 이해가 선행되어야 한다.
- 요약된 값이 무엇을 뜻하는지, 각 항목의 의의를 알지 못하면 데이터를 올바르게 파악하기 어렵다.



▶ 정형 데이터와 비정형 데이터



- 정형 데이터는 미리 정해 놓은, 잘 알려진 포맷이나 명확한 데이터 구조 표현 방법을 사용하므로 사용자가 쉽게 이해하고 시스템에 적용할 수 있다.
- 정형화된 업무 또는 서비스 등에 활용한다.
- 스프레드시트, 관계형 데이터베이스의 테이블, CSV 등

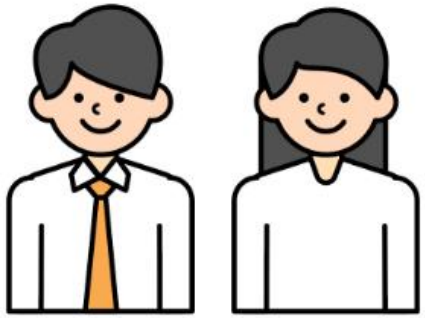
- 반정형 데이터는 정형 데이터처럼 테이블(table)로 구조화되어 있지는 않으나, 파일에 포함된 데이터 구조 정보를 바탕으로 데이터를 매핑(mapping)하여 정형 데이터로 변환할 수 있다.
- HTML, XML, JSON, RDF, 로그(Log) 데이터, 센싱(Sensing) 데이터 등

- 비정형 데이터는 사전에 정의된 형식과 구조가 없는 데이터로서 다양하고 방대한 양의 데이터이므로 별도의 분석 처리 기술이 필요하다.
- 텍스트, 이미지, 음원 데이터, 빅 데이터 등

▶ 정량적(Quantitative) & 정성적(Qualitative) 데이터

분류	척도	설명
범주형 (Categorical)	명목 변수 (Nominal)	• 범주/그룹에 배치할 수 있는 데이터를 나타내는 변수로 범주 자체는 고유한 순서나 비교가 불가능함 • 명목 변수는 질적 특성을 가지며 성별, 인종, 종교와 같은 집단 특성을 나타낼 수 있음 • 성별, 결혼 상태, 머리색 등 • 예) {서울, 대전, 대구, 제주}
	서열 변수 (Ordinal)	• 범주/그룹에 배치 가능한 범주형 변수이지만 자연스러운 순서/순위가 존재함 • 범주에 대한 상대적 가치나 중요도 측면에서 순위를 매기거나 정렬이 가능하지만 간격이나 비율이 의미 없음 • 등수(1등, 2등, 3등), 경제적 지위(저소득, 중간소득, 고소득), 리커트 척도 등 • 예) {매우좋음, 좋음, 보통}
수치형 (Numerical)	연속형 변수 (Continuous)	• 특정 범위 내에서 모든 값을 가질 수 있는 수치 변수 • 연속체를 따라 어느 지점에서나 측정할 수 있으며, 종종 시간, 거리 및 온도와 같은 측정 단위를 포함할 수 있음 • 높이, 온도, 키, 무게, 농도, 소득 등 소수점을 포함하는 값 • 예) 전국민의 키 측정 자료
	이산형 변수 (Discrete)	• 특정 범위 내에서의 특정 값만을 취할 수 있는 수치 변수로 개수 또는 정수가 포함됨 • 일반적으로 데이터 세트의 범주나 그룹을 나타내는 데 사용됨 • 자녀 수, 직원 수, 구매한 항목 수 등 소수점을 포함할 수 없는 정수 • 예) 각 부서의 직원 수

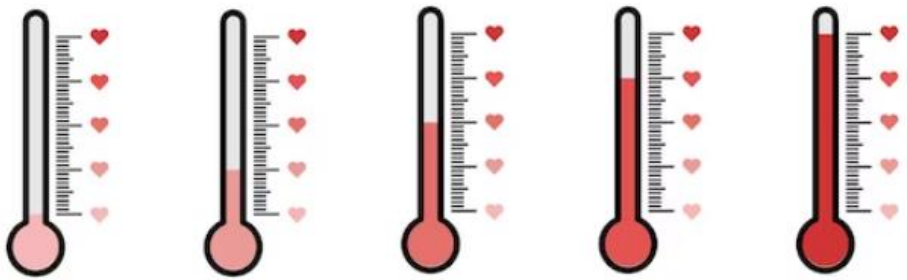
▶ 데이터는 대상에 대한 측정 가능한 표현이다



범주형 – 명목형 데이터
"남성", "여성"

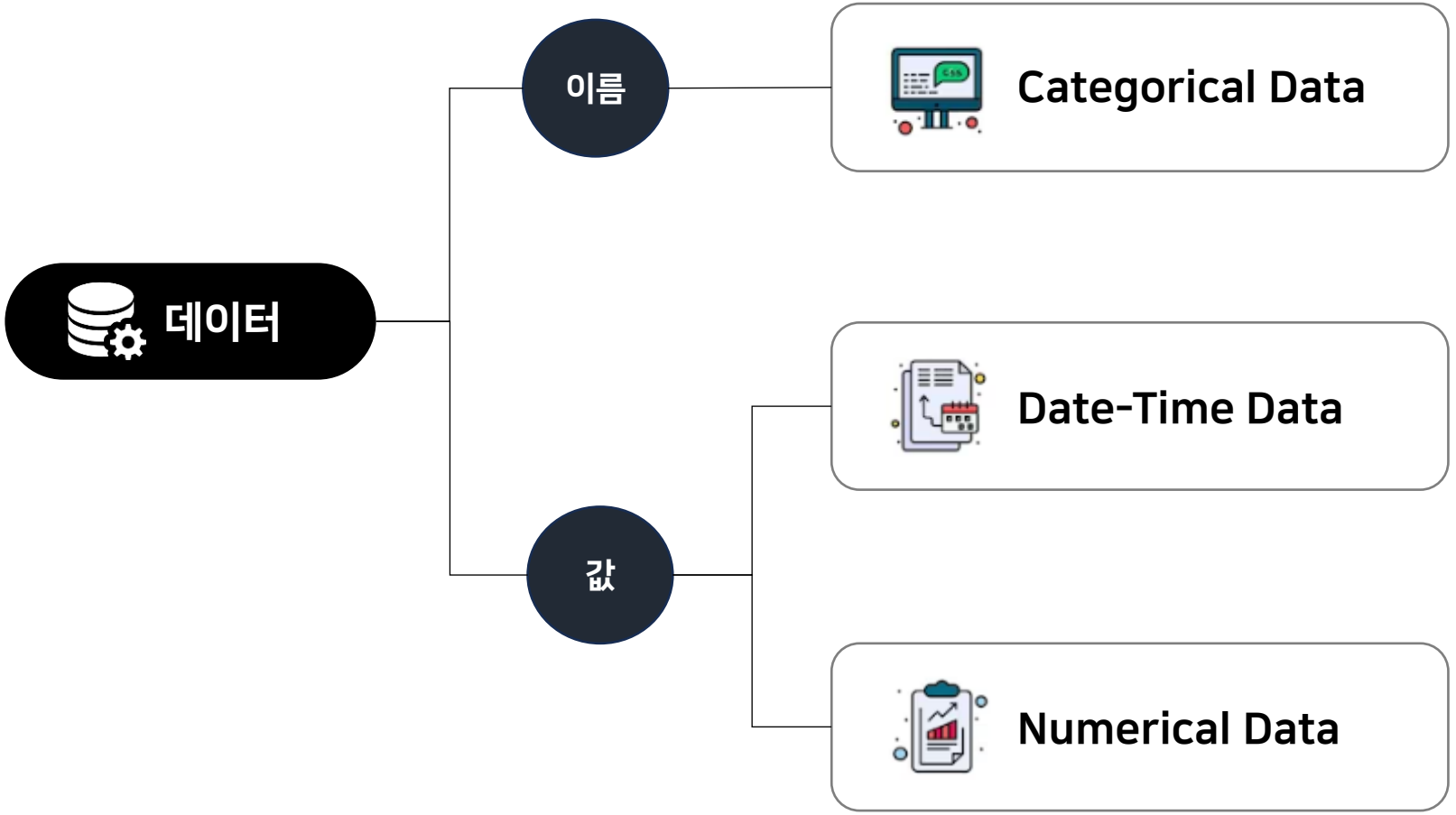
범주형 – 서열형 데이터
1등, 2등, 3등

수치형 – 이산형 데이터
1, 2, 3, 4, 5



수치형 – 연속형 데이터
21.2, 36.5, 40.0

▶ 엑셀의 기능적 데이터 유형 구분



▶ 같은 결과를 내는 3가지 방법

MFC사의 47,646건 수익 자료의 KPI 요약표이다. 빈 칸에 알맞은 값은?

통계량	매출	이익	이익률
평균	㉠	1,722,456	44.63%
최소값	2,686,000	927,372	31.98%
최대값	4,999,000	3,807,640	62.55%
표준편차	668,369	㉡	3.95%

① ㉠ 3,844,503 ㉡ 338,135
② ㉠ 3,844,503 ㉡ 1,038,135
③ ㉠ 4,344,503 ㉡ 338,135
④ ㉠ 4,344,503 ㉡ 1,038,135

▶ 엑셀 컴포넌트 한 눈에 파악하기

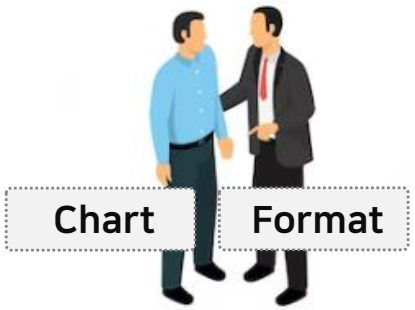
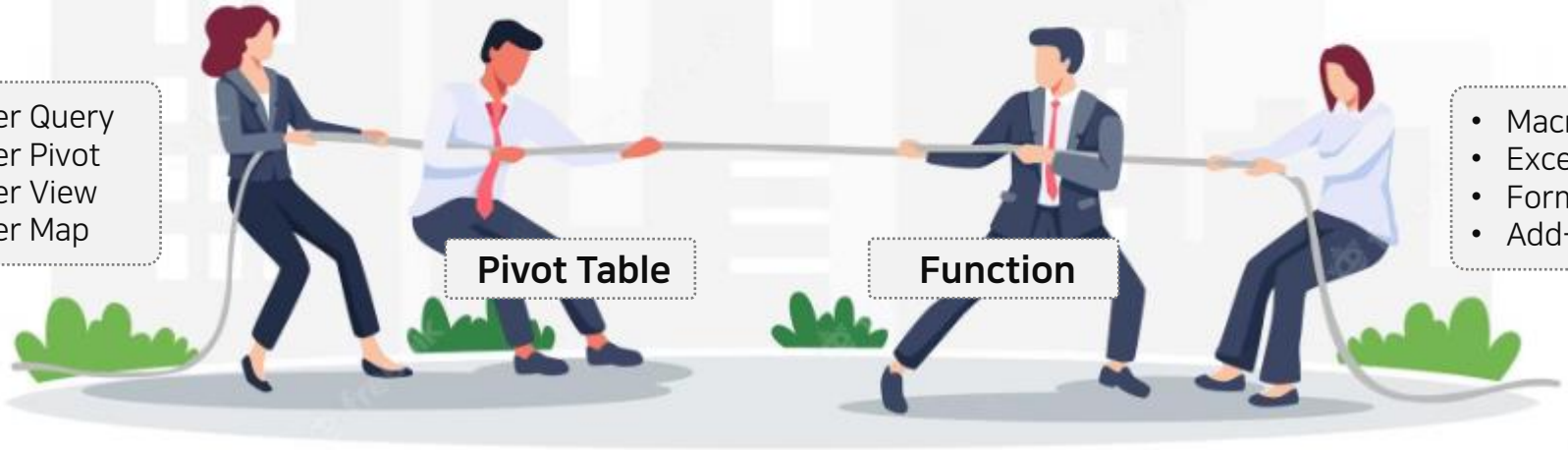


Chart Format

- Calcurators
- Solver
- Forecast Sheet
- Senario
- Data Table
- Filter
- ...

- Power Query
- Power Pivot
- Power View
- Power Map



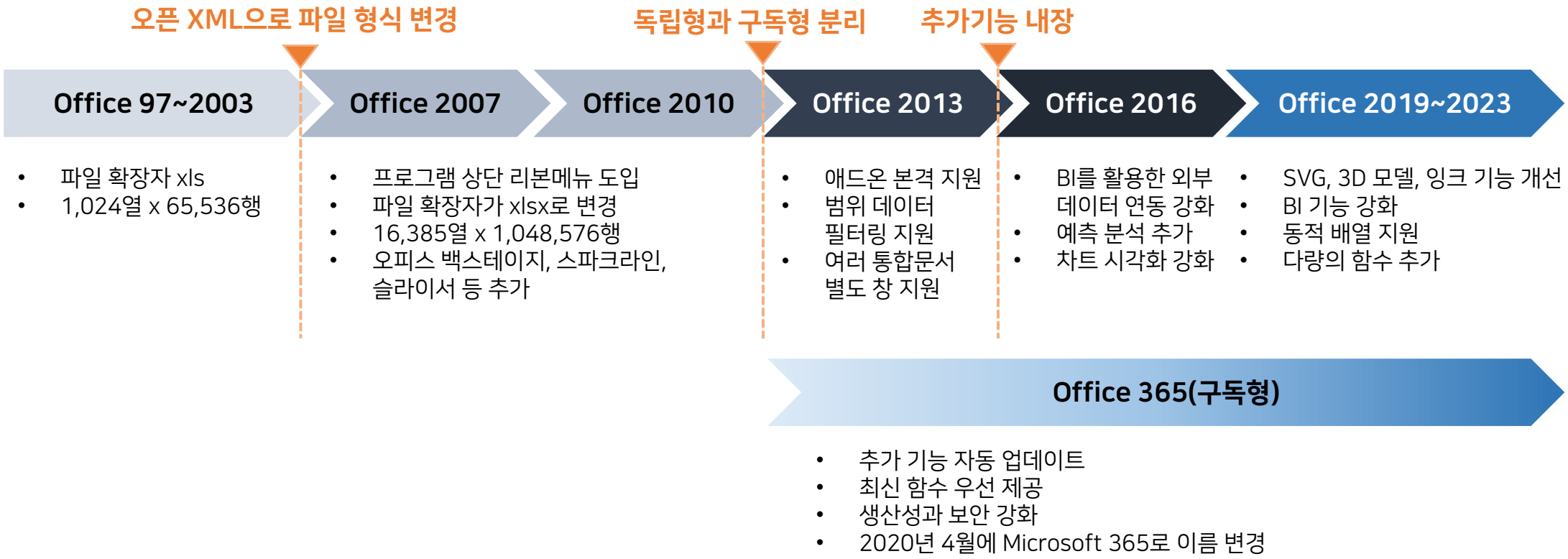
Pivot Table

Function

- Macro
- Excel VBA
- Form Control
- Add-on

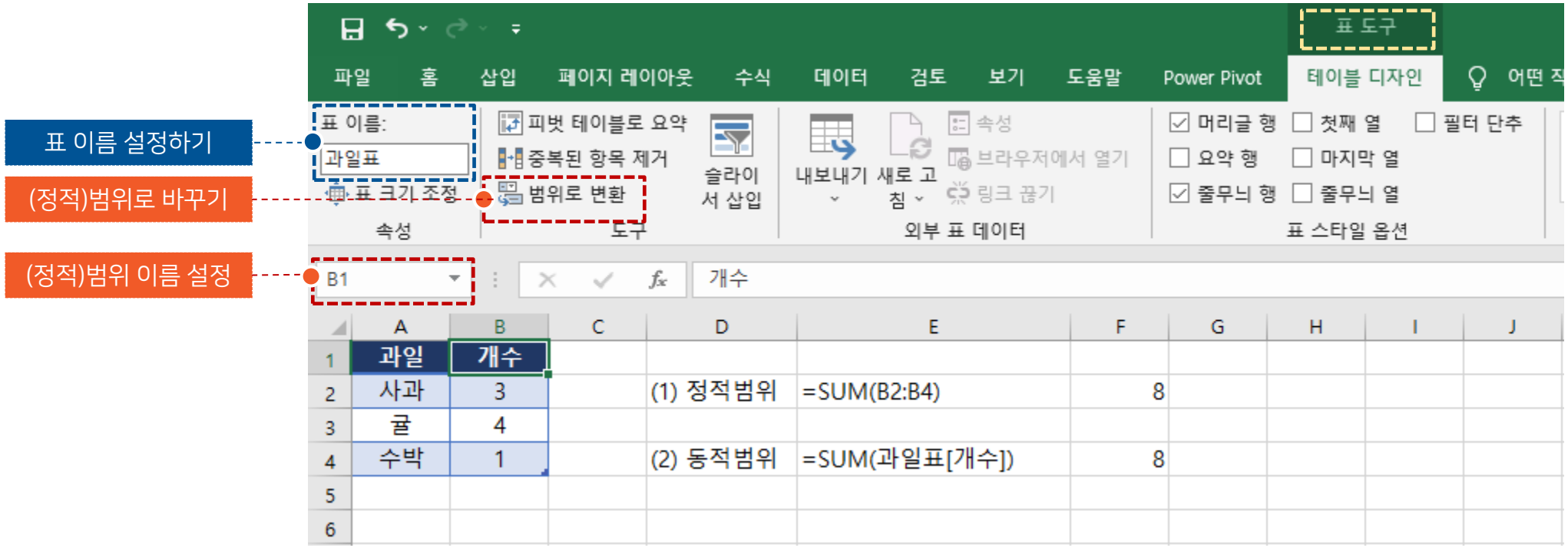
EXCEL BI ⟷ **함수 + VBA**

▶ 엑셀 버전 정리


오픈 XML으로 파일 형식 변경
독립형과 구독형 분리
추가기능 내장
Office 97~2003
Office 2007
Office 2010
Office 2013
Office 2016
Office 2019~2023
• 파일 확장자 xls
• 1,024열 x 65,536행
• 프로그램 상단 리본메뉴 도입
• 파일 확장자가 xlsx로 변경
• 16,385열 x 1,048,576행
• 오피스 백스테이지, 스파크라인, 슬라이서 등 추가
• 애드온 본격 지원
• 범위 데이터 필터링 지원
• 여러 통합문서 별도 창 지원
• BI를 활용한 외부 데이터 연동 강화
• 예측 분석 추가
• 차트 시각화 강화
• SVG, 3D 모델, 잉크 기능 개선
• BI 기능 강화
• 동적 배열 지원
• 다량의 함수 추가
Office 365(구독형)
• 추가 기능 자동 업데이트
• 최신 함수 우선 제공
• 생산성과 보안 강화
• 2020년 4월에 Microsoft 365로 이름 변경

▶ 동적 범위와 정적 범위

- **엑셀의 표(동적범위)와 범위(정적범위) 중 "표"로 설정하여 데이터를 관리한다.**
- 새로운 데이터가 추가되면 동적범위(표)로 계산한 수식은 자동으로 데이터가 확장되지만, 정적범위는 변화가 없다.

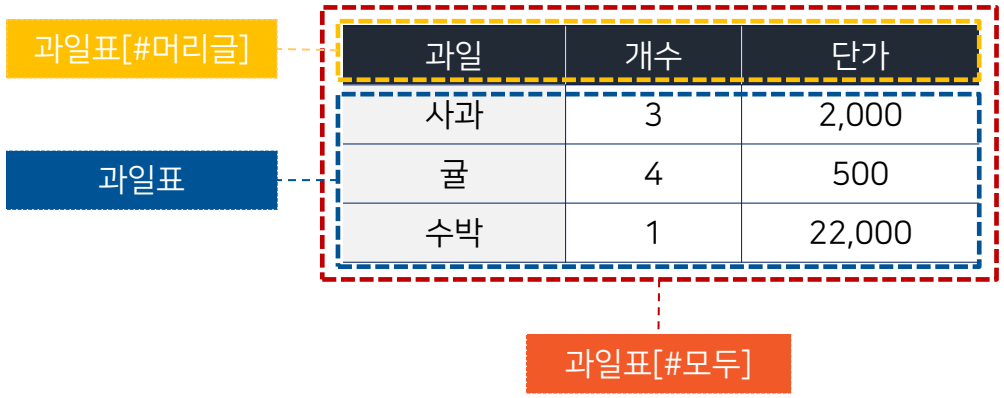


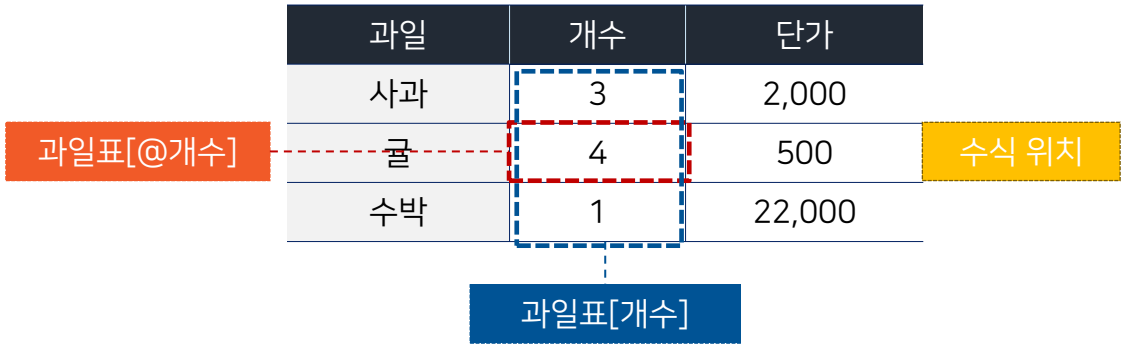
▶ 표의 구조적 참조

- 표의 구조적 참조란, 표 이름과 필드명에 간단한 기호를 추가해 수식을 작성하는 표의 고유 기능이다.
- 구조적 참조 명령어

구조적 참조 명령어	참조하는 내용
표이름	표의 데이터 중 머리글을 제외한 나머지 부분(2행~마지막 행까지)을 모두 참조한다
표이름[#모두]	머리글을 포함한 표의 모든 범위(1행~마지막 행까지)를 참조한다
표이름[#머리글]	표의 머리글에 해당하는 범위(1행)만 참조한다
표이름[필드명]	표의 특정 필드에 해당하는 모든 데이터를 참조한다
표이름[@필드명]	해당 필드의 여러 데이터 중 수식이 입력된 셀과 같은 행에 들어 있는 값을 참조한다

- "과일표"의 구조적 참조 예시





▶ 분석 데이터의 관리 형식

- 엑셀 자료의 일반적 데이터 관리 구조는 테이블(table), 크로스탭(crosstab), 템플릿(template)의 3가지이다.
- 원활한 데이터 분석을 위해서는 데이터 형식은 테이블 형태로 구성되어야 한다.

회사 이름

나머지 주소
시/도, 우편 번호

전화: 전화 번호
팩스: 팩스 번호

전자 메일
웹 사이트

청구 대상: 콘토소(주)
주소: 부평구 부평동
인천, 인천 098765

전화 번호: 432-555-0189
팩스: 432-555-0123
전자 메일: someone@example.com

송장 번호: 3-456-2
송장 날짜: 날짜

송장 대상: 프로젝트 2

항목 번호	설명	수량	단가	할인	가격
Z4567	송장 3-456-2 데이터 1	39	₩ 5.00	₩ -	₩ 195.00
Z4568	송장 3-456-2 데이터 2	40	₩ 4.00	₩ 5.00	₩ 155.00
Z4569	송장 3-456-2 데이터 3	30	₩ 6.00	₩ 7.00	₩ 173.00
Z4570	송장 3-456-2 데이터 4	40	₩ 7.00	₩ -	₩ 280.00
Z4571	송장 3-456-2 데이터 5	10	₩ 4.00	₩ -	₩ 40.00
Z4572	송장 3-456-2 데이터 6	5	₩ 8.00	₩ -	₩ 40.00
Z4573	송장 3-456-2 데이터 7	70	₩ 6.00	₩ -	₩ 420.00
Z4574	송장 3-456-2 데이터 8	25	₩ 4.00	₩ -	₩ 100.00

[템플릿(template)]

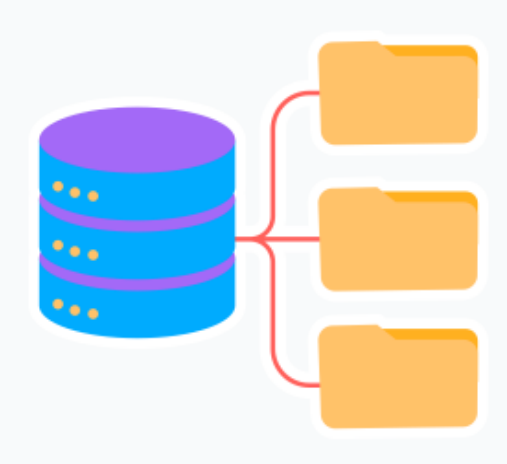
마켓분류	거래량	매출액	수익	배송비
LATAM	10,294	2,780,430,284	2,191,098,898	301,885,283
USCA	10,378	3,036,718,603	2,304,508,724	317,853,915
아시아 태평양	14,302	5,192,787,354	4,034,562,184	562,421,642
아프리카	4,587	1,006,754,423	837,558,786	114,112,034
유럽	11,729	4,222,577,617	3,397,719,295	448,162,735

[테이블(table) 형식]

마켓분류	1월	2월	3월	4월
LATAM	341	277	578	723
USCA	414	338	702	672
아시아 태평양	861	684	772	731
아프리카	318	244	278	274
유럽	686	646	696	644

[크로스탭(crosstab) 형식]

▶ 테이블(table)의 조건



① 테이블은 1개 행으로 이뤄진 머릿글을 갖는다.

② 테이블의 데이터는 열(column) 방향으로 추가하고 관리된다.

③ 1개 열에는 같은 속성(attribute) 데이터만 들어 있다.

④ 같은 속성(attribute) 데이터는 1개 열에만 들어 있다.

▶ 테이블의 필드와 레코드

- **데이터로부터 의미있는 인사이트를 도출하려면 체계적이며 상세히 정리된 열(변수 =필드)이 필요하다.**
- 데이터 세트의 열(column)은 필드(filed), 변량(variate), 변수(variable), 속성(attribute), 특성(feature), 차원(dimension)으로도 불리며, 데이터 분석의 관점을 나타낸다.

- 열 이름(필드명)은 누구나 알아보기 쉽게 작성해야 한다. 불가피하게 복잡해진다면 참조할 데이터 정의 기술서가 필요하다.
- 의미를 제대로 파악할 수 없는 데이터를 활용한 데이터 분석으로 얻는 결론은 신뢰하기 어렵다.
  여러 사람이 데이터를 공동 관리하면 의미를 알 수 없는 필드가 발생하기 쉬우므로 꾸준한 관리가 필요하다.

▶ 데이터 분석 프로세스

데이터 수집
- 적정 규모의 샘플 데이터 표집
- 다양한 취합 방식
  (crawling, survey, tagging, logging 등)

데이터 전처리
- 데이터 유형, 척도 등 정리
- 결측값 처리
- 테이블(table) 형태 변환
- 표준화와 정규화
- …

Clean Dataset

탐색적 데이터 분석
EDA, Exploratory Data Analysis
- 기술통계법
- 일변량 데이터 탐색
- 다변량 데이터 탐색
- 이상치 탐색
- …

분석 모델 & 알고리즘 적용
- 회귀 예측 모델
- 분류와 군집화
- 시계열 분석
- 지도, 비지도, 강화 학습
- 이미지, 감성 분석
- 시뮬레이션
- …

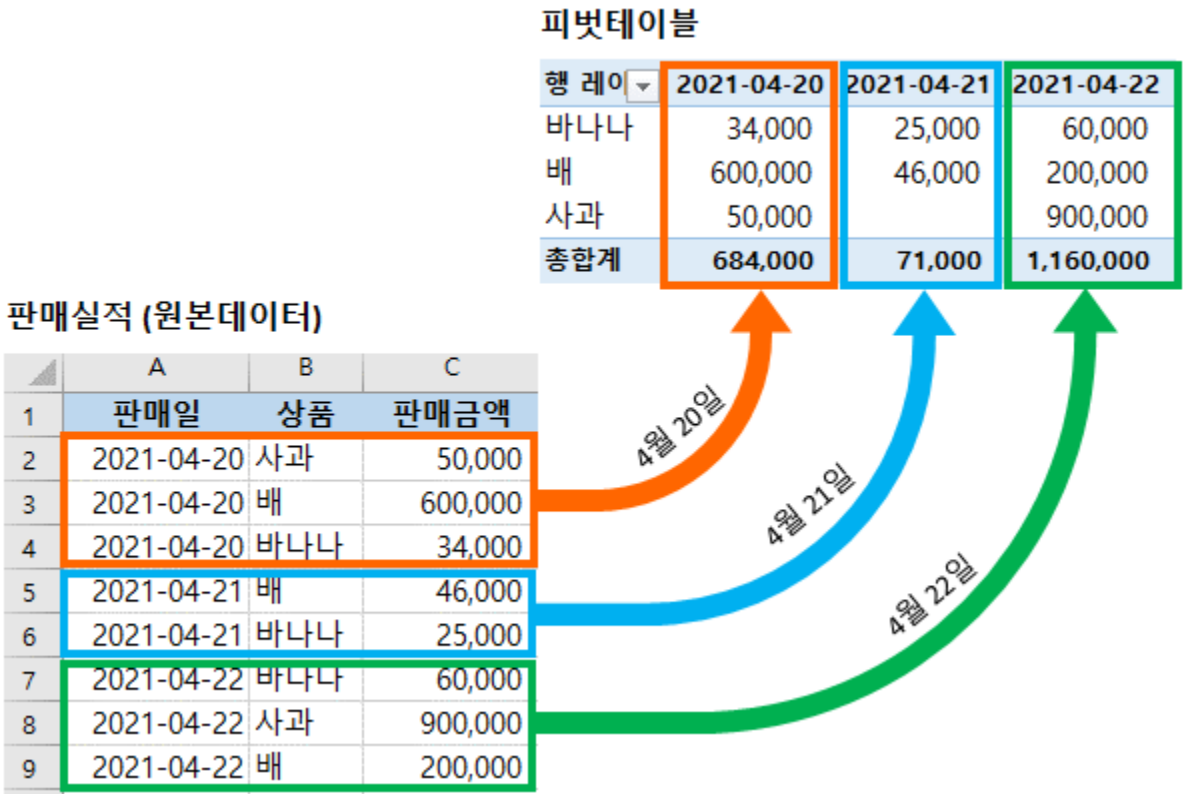
데이터 기반 커뮤니케이션
- 데이터 분석 결과의 시각화 보고서를 활용한 데이터 기반 커뮤니케이션
- …

Decision Making

현실세계
Business Question

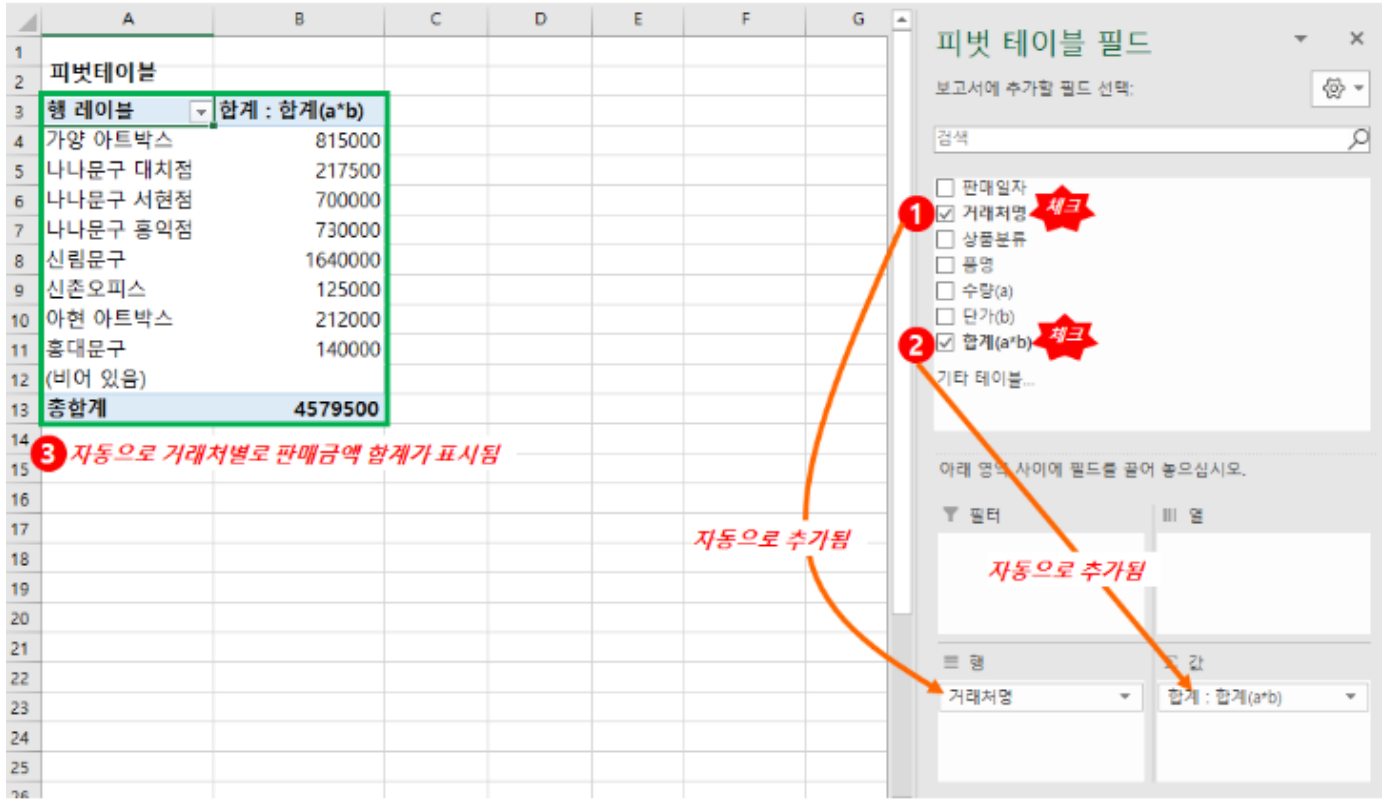
▶ 피벗테이블(pivot table)

- Pivot : 회전하는 물체의 균형을 잡아 주는 중심점. 쉽게는 원의 중심정 정도로 생각할 수 있음
- 엑셀 피벗테이블은 자신이 원하는 데이터를 원하는 행과 열에 데이터를 배치해서 새로운 집계값을 만드는 기능
- 피벗테이블은 워크시트에 입력된 많은 양의 데이터에서 필요한 자료만을 뽑아 새롭게 집계표를 작성해주는 기능으로,
  피벗테이블을 사용하면 엑셀을 활용한 데이터분석을 매우 효과적으로 진행할 수 있도록 해주는 엑셀 데이터분석의 핵심 기능



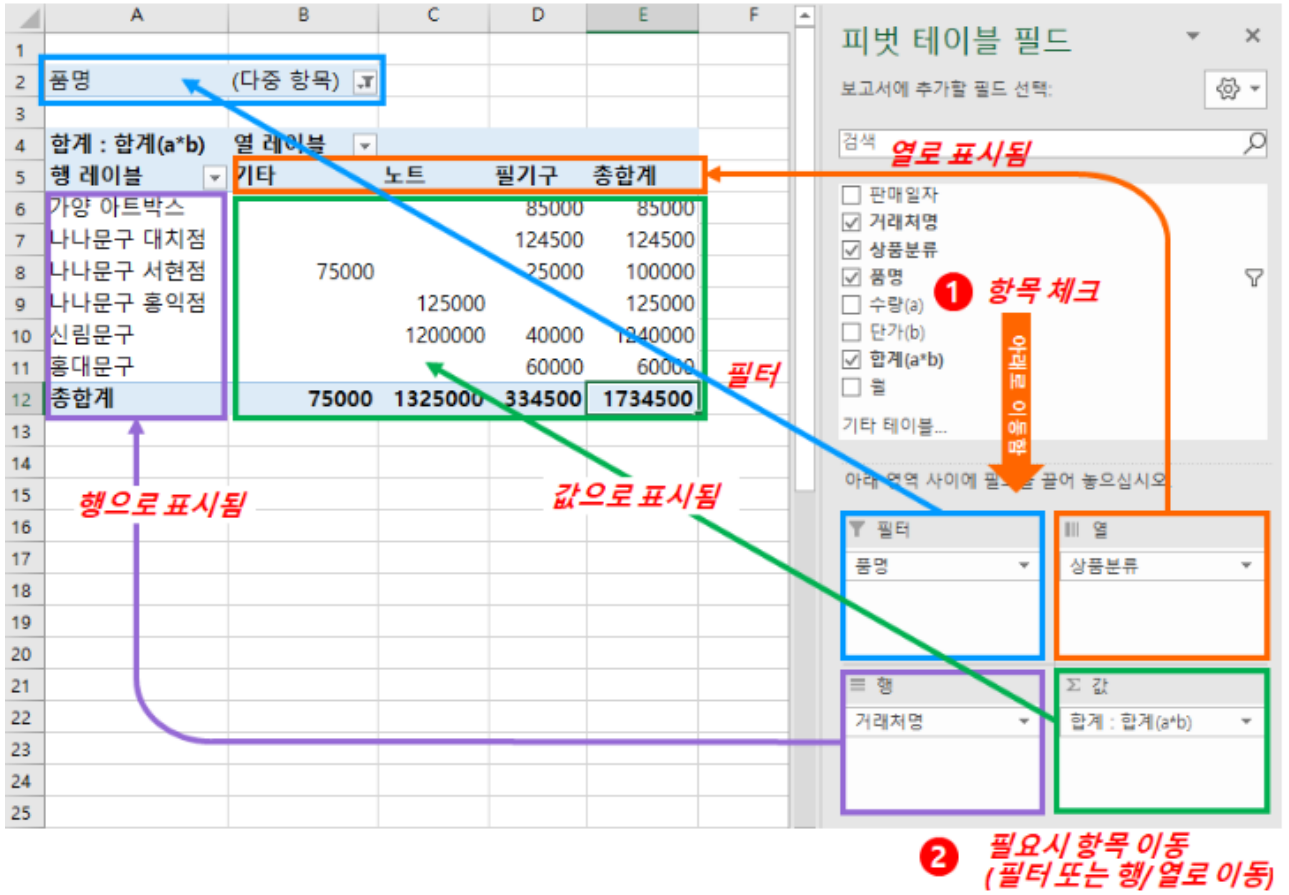
▶ 피벗테이블 보고서 생성

- 필드 목록에서 원하는 피벗테이블 필드를 체크하면, 필드 영역에 표시되면서 보고서가 만들어진다.
- 범주형(Categorical) 데이터 – 기본적으로 “행” 영역으로 이동하며, 행에 놓인 필드는 다른 영역으로 이동 가능하다.
- 수치형(Numeric) 데이터 – 기본적으로 “값” 영역으로 이동하여 “합계”를 기본으로 표시한다.
- 날짜(Datetime) 데이터 – 기본적으로 “행” 영역으로 이동하며, 버전에 따라 자동 그룹화된다.



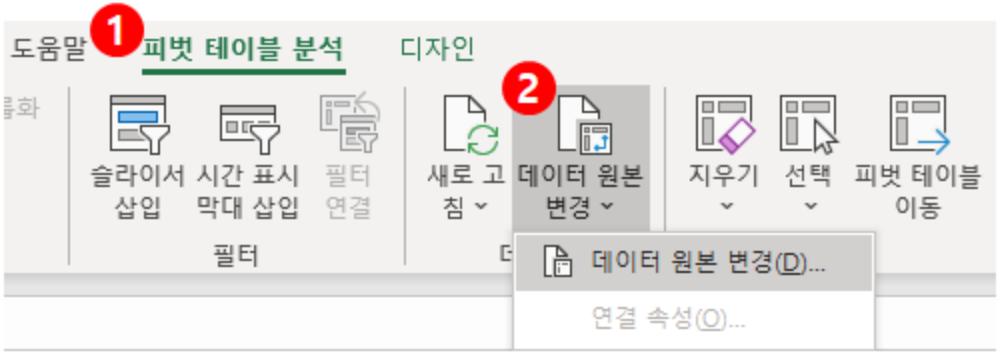
▶ 피벗테이블 보고서 다루기

- 행 영역의 필드를 열 영역으로 이동하면 필드가 열 단위로 표시된다.
- 수치형이 아닌 데이터(범주형, 날짜형)를 값 영역으로 이동하면, 값 요약 기준이 '합계'가 아닌 '개수'가 된다.
- 필드를 필터 영역으로 이동하면 조회 조건으로 사용된다.



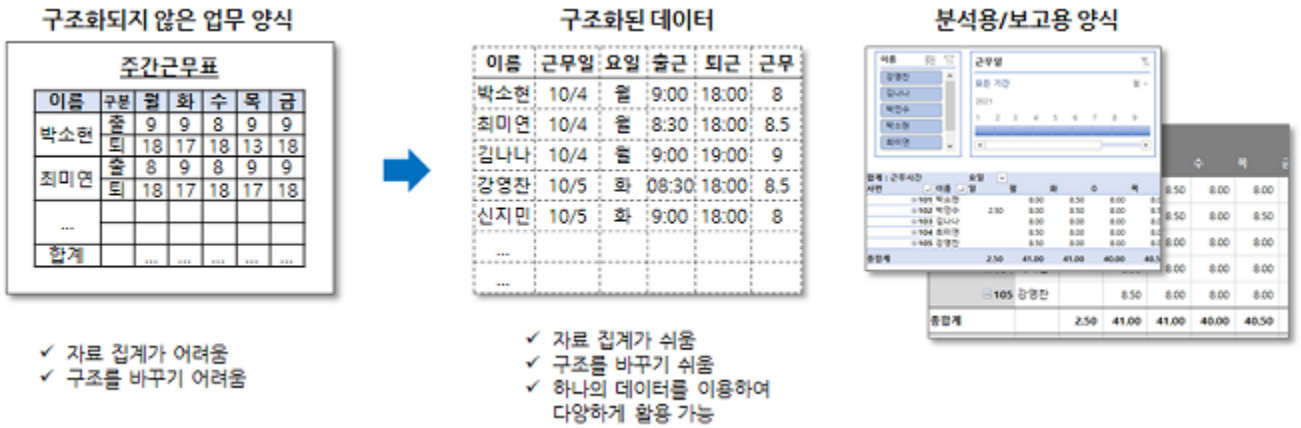
▶ 피벗테이블 참조 범위 새로고침

- 원본 데이터가 변경되거나 새로운 행이 추가되어도 피벗테이블의 값은 바뀌지 않으므로 [새로 고침] 한다.
- 범위를 선택하여 피벗테이블을 만든 후, 이 범위를 벗어난 데이터가 추가되면 피벗테이블에서 [새로 고침]해도 변경되지 않으므로 [피벗테이블 분석] > [데이터 원본 변경에서 범위를 변경해줘야 한다.



▶ 데이터와 분석/보고용 양식을 분리하기

- 엑셀로 자료를 잘 다루려면 데이터는 구조화된 데이터 양식(동적 테이블)로 관리하고,
  분석용 또는 보고용 양식은 별도의 템플릿을 활용하여 분리해야 한다.

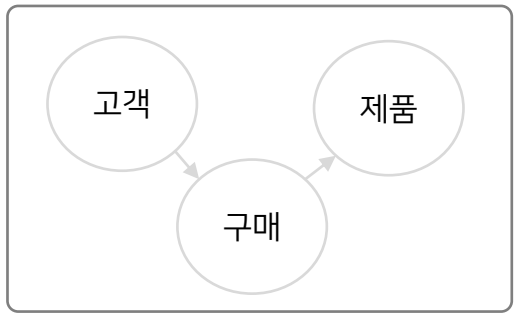


- 데이터가 구조화되어 있어야 엑셀의 기능(피벗테이블, 필터, 정렬 등)과 함수, 수식을 쉽게 사용할 수 있다.
- 구조화된 데이터로부터 다양한 형태의 분석 및 보고용 양식을 손쉽게 만들 수 있다.
- 업무에 필요한 항목이 있을 때 필드(열)를 손쉽게 추가할 수 있으며 기존의 데이터와 기존의 수식에 영향을 미치지 않는다.

▶ 현실 세계의 추상화

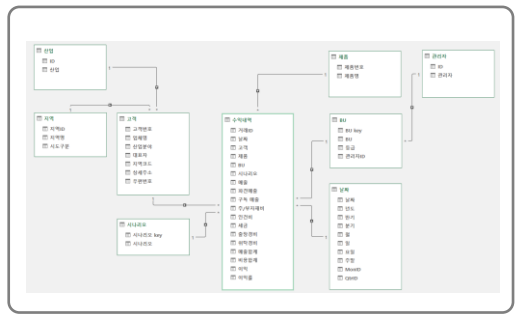
현실 세계

- 데이터가 나타내고자 하는 실제 세계의 현상



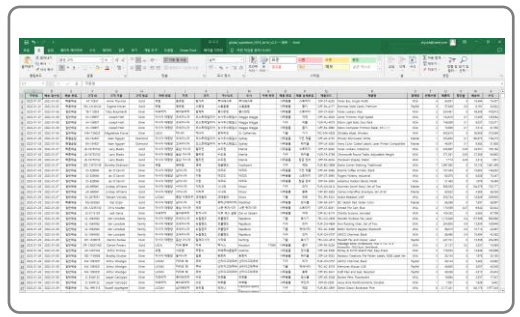
개념적 데이터 모델

- 도메인 모델(Domain Model)
- 시스템이 포함할 내용과 구성, 비즈니스 규칙을 거시적으로 정리한 추상적 모델
- 비즈니스에 중요하고 데이터 모델에 반영되어야할 요소(Entity Class) 정의, 엔터티 클래스의 특징, 제약조건, 관계, 보안 및 무결성 요구 등을 포함한다.



논리적 데이터 모델

- 도메인의 개념 및 관계에 관한 구체적 정보
- 데이터 유형과 길이 등 데이터 속성, 엔터티 간의 관계 등을 표현한다.

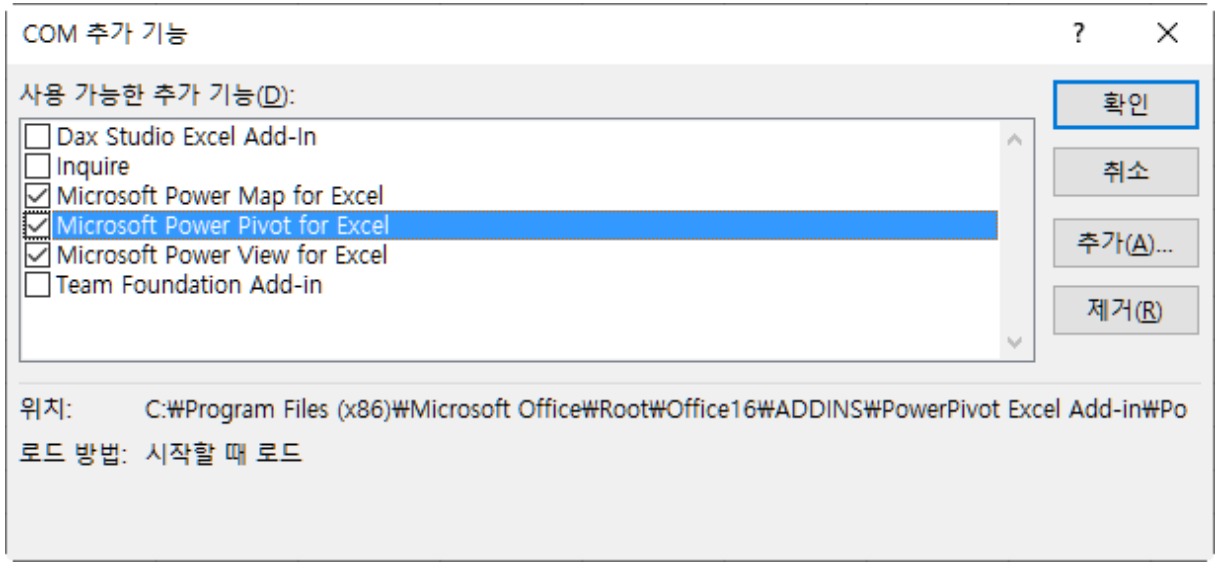


물리적 데이터 모델

- 데이터베이스 내에서 데이터가 물리적으로 저장되는 방식

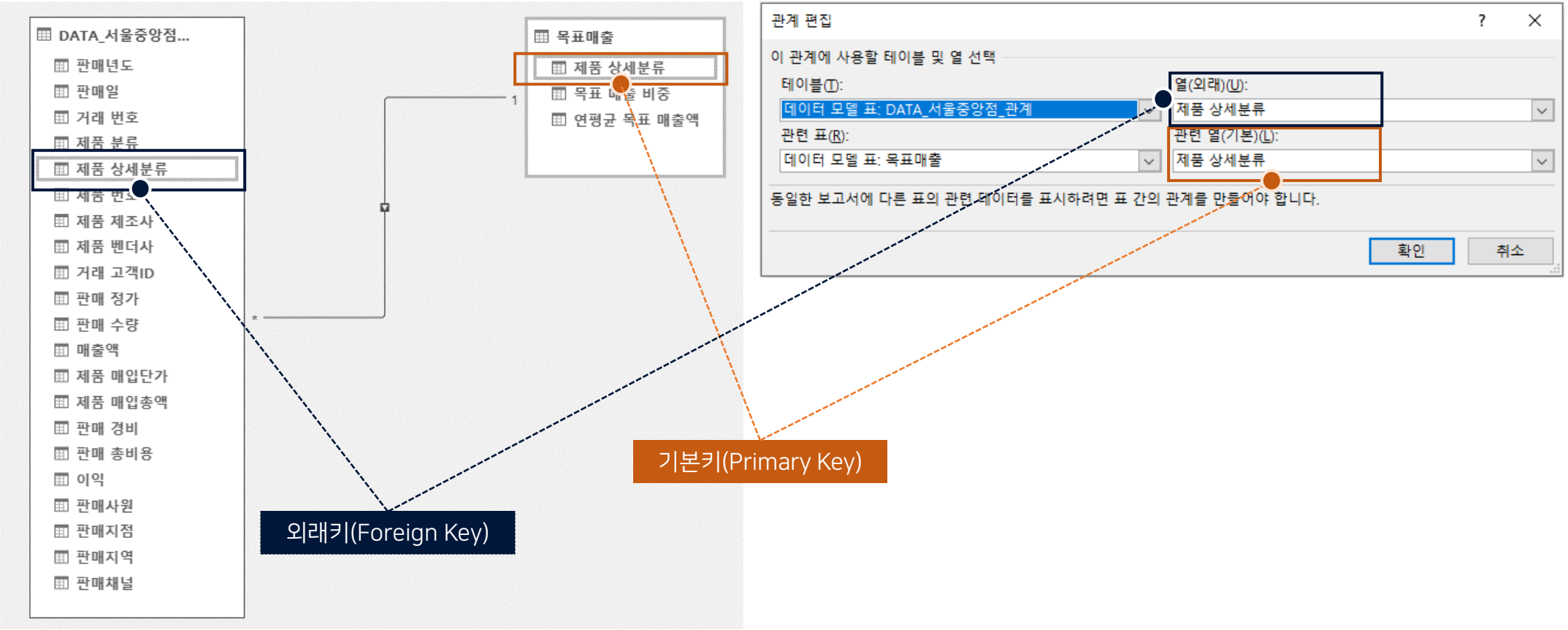
▶ Power Pivot 추가 기능 로드

- [개발 도구] – [추가 기능] – [COM 추가 기능] – [Microsoft Power Pivot for Excel]



▶ 데이터 모델과 키

- **데이터 모델(Data Model)**이란 현실 세계의 정보 구조를 컴퓨터로 처리하기 위해 단순화, 추상화해 체계적으로 표현한 개념 모형이다.
- 물리적으로 분리된 두 테이블이 가진 공통 필드를 **키(Key)**라 한다.
  유일값(Unique value)을 갖는 쪽을 **기본키(Primary Key)**라 하며, 여러 번 반복 등장하는 쪽을 **외래키(Foreign Key)**라 한다.



▶ 데이터 모델

- 데이터베이스를 구축할 때 체계화된 구조를 갖추기 위해 필요한 개념들의 집합(테이블, 필드, 관계, 기본키, 외래키 등)
- 이러한 구조에서 필요한 연산, 구조와 연산에 대한 제약 조건 등을 포함

▶ 데이터 모델링

- 데이터 모델을 설계하는 것
- 데이터 모델을 사용하면 여러 테이블의 데이터를 통합하여 Excel 통합 문서 내에서 관계형 데이터 원본을 효율적으로 작성
- 테이블 설계, 관계 설정, DAX(Data Analysis Expression)를 활용한 분석식 작성

▶ DAX (Data Analysis eXpression)

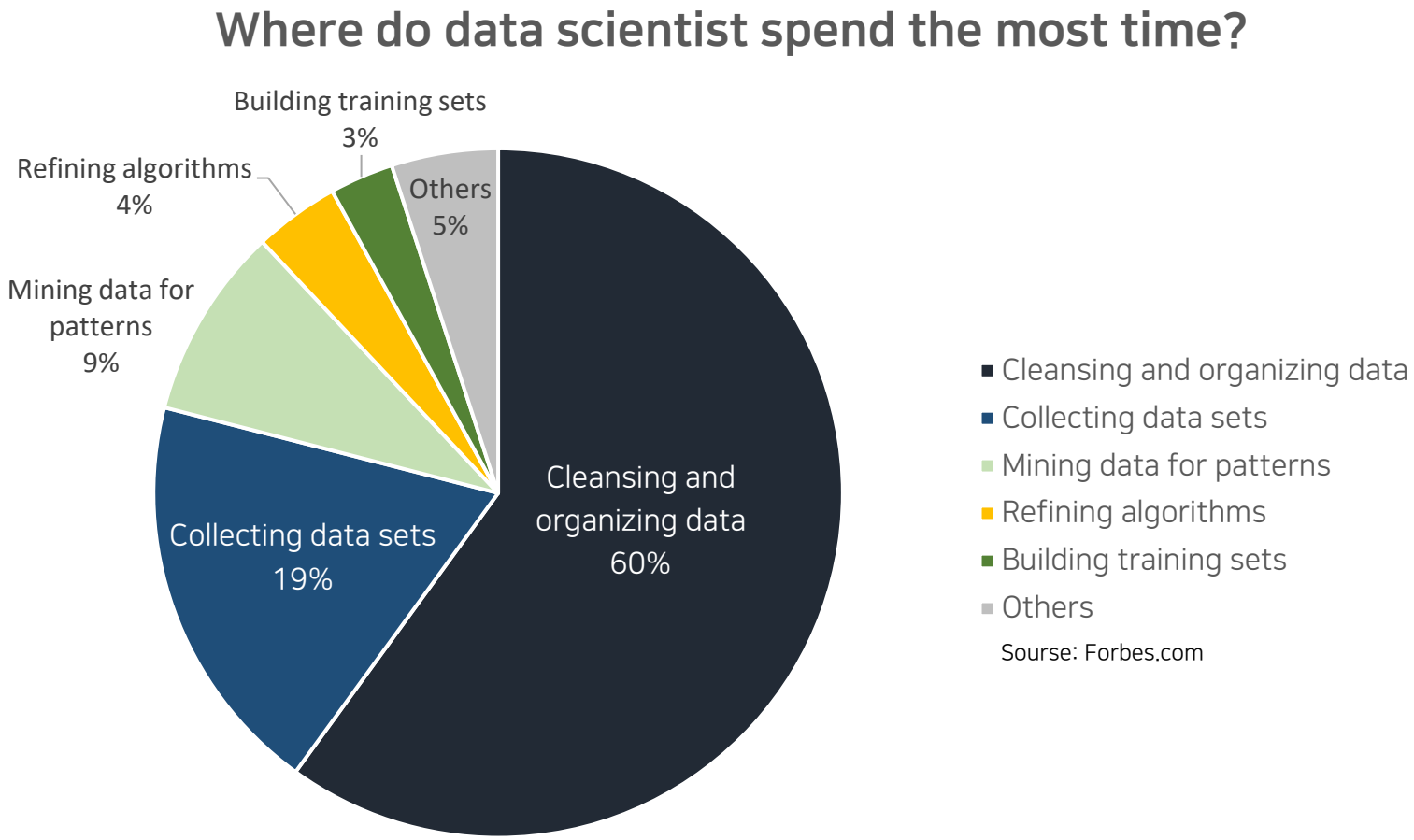
- DAX를 사용하면 모델에 있는 데이터를 사용하여 새로운 정보를 만들 수 있음
- Excel 함수와 유사하지만 더 복잡한 분석 함수 제공
- 셀 단위 참조가 아닌 필드(열), 레코드(행) 단위로 동작
- = '테이블명' [필드명] (테이블명 생략 가능)

▶ DAX 수식의 종류

- **계산 열**
  - 파워 피벗 테이블에 열을 추가하여 계산
  - 테이블의 각 행에 대해 계산
- **측정값**
  - 피벗 테이블과 피벗 차트에서 사용하기 위해 작성되는 수식
  - 피벗 테이블의 값 영역에 사용
  - 집계 함수를 사용한 집계 값 또는 고급 계산

▶ 데이터 전처리의 중요성

- 수식이 복잡하고 어려워지는 이유는 대체로 자료가 구조화되지 않았기 때문에 문제가 발생한 것이다.



▶ 전처리가 필요한 엑셀 자료

Quiz

다음 구조화되지 않은 엑셀 자료의 문제점은 무엇인가?

거래처별 판매 이력

판매일자	거래처	상품		수량(a)	단가(b)	합계(a*b)
		분류	품명			
2021-01-02	나나문구 홍익점(직영대리점)	기타	포스트잇 노트 (654) 노랑	70	1,700	119,000
2021-01-02	나나문구 홍익점(직영대리점)	기타	포스트잇 노트 큐브 3색	86	2,300	197,800
2021-01-02	나나문구 홍익점(직영대리점)	복사용지	더블에이 A4용지	42	20,000	840,000
2021-01-02	신림문구(가맹대리점)	기타	오피스 수정테이프	0	20,000	0
2021-01-02	신림문구(가맹대리점)	노트	옥스포드 노트	104	6,000	624,000
2021-01-02	신림문구(가맹대리점)	노트	카카오프렌즈 인덱스 노트 네오	33	5,000	165,000
2021-01-02	신림문구(가맹대리점)	필기구	모나미 볼펜	90	100	9,000
						0
2021-01-03	가양 아트박스(직영대리점)	노트	합지 스프링노트	121	2,500	302,500
2021-01-04	신촌오피스(직영대리점)	노트	카카오프렌즈 인덱스 노트 라이	28	5,000	140,000
2021-01-05	나나문구 서현점(직영대리점)	기타	데스크 오거나이저		15,000	0
2021-01-05	나나문구 서현점(직영대리점)	복사용지	더블에이 A4용지	60	20,000	1,200,000
2021-01-05	나나문구 서현점(직영대리점)	필기구	모나미 볼펜	133	100	13,300

▶ 전처리가 필요한 엑셀 자료

앞 장의 자료에 전처리가 필요한 이유

- 실제 자료는 4행부터 시작하므로 정렬, 필터 등을 사용하려면 항상 범위를 지정해 주어야 한다.
- 데이터를 정렬하기가 불편하다. 정렬을 할 수 있긴 하지만 머리글을 제대로 못 가져오는 문제가 있다.
- 머리글이 병합되어 있어서 일부 필드는 필터기능을 사용할 수 없다.
- 머리글이 병합되어 있어서 피벗테이블을 사용할 수 없다.
- 머리글이 병합되어 있어서 일부 필드는 부분합 기능을 사용할 수 없다.
- 거래처는 두 가지 정보가 포함되어 있어서 제대로 사용할 수 없다. 텍스트 나누기를 해야 한다.
- 빈 행이 중간에 포함되어 있어서 빈 행을 계산에 포함할 때 일부 계산이 잘못될 수 있다.

	A	B	C	D	E	F	G	H
1	판매일자	거래처명	대리점유형	상품분류	품명	수량(a)	단가(b)	합계(a*b)
2	2021-01-02	나나문구 홍익점	직영대리점	기타	포스트잇 노트 (654) 노랑	70	1,700	119,000
3	2021-01-02	나나문구 홍익점	직영대리점	기타	포스트잇 노트 큐브 3색	86	2,300	197,800
4	2021-01-02	나나문구 홍익점	직영대리점	복사용지	더블에이 A4용지	42	20,000	840,000
5	2021-01-02	신림문구	가맹대리점	기타	오피스 수정테이프	53	20,000	1,060,000
6	2021-01-02	신림문구	가맹대리점	노트	옥스포드 노트	104	6,000	624,000
7	2021-01-02	신림문구	가맹대리점	노트	카카오프렌즈 인덱스 노트 네오	33	5,000	165,000
8	2021-01-02	신림문구	가맹대리점	필기구	모나미 볼펜	90	100	9,000
9	2021-01-03	가양 아트박스	직영대리점	노트	합지 스프링노트	121	2,500	302,500
10	2021-01-04	신촌오피스	직영대리점	노트	카카오프렌즈 인덱스 노트 라이	28	5,000	140,000

▶ 전처리가 필요한 엑셀 자료

Quiz

인쇄된 보고서 형식의 자료를 엑셀에 그대로 옮겨 주(week) 단위로 시트를 만드는 자료가 있다.
건설현장 작업일보, 생산현장 생산일보 등 비슷한 모습의 엑셀 자료가 가진 문제점은 무엇인가?

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		주간 근무표										
3												
4		사번	이름	구분	10월 04일	10월 05일	10월 06일	10월 07일	10월 08일	10월 09일	10월 10일	합계
5					월	화	수	목	금	토	일	
6		101	박소현	출근	8:00	9:00	9:00	9:00	8:30	9:00		47.50
7				퇴근	17:00	18:00	18:00	18:00	17:00	12:00		
8		102	박민수	출근	8:00	9:00	9:00	9:00	8:20		9:00	48.17
9				퇴근	17:00	18:00	18:00	18:00	17:00		12:30	
10		103	김나나	출근	8:00	8:00	8:00	8:00	8:00			45.00
11				퇴근	17:00	17:00	17:00	17:00	17:00			
12		104	최미연	출근	9:00	9:00	9:00	9:00	9:00			45.00
13				퇴근	18:00	18:00	18:00	18:00	18:00			
14		105	강영찬	출근	9:00	8:00	8:00	8:00	9:30	9:00		46.00
15				퇴근	18:00	17:00	17:00	17:00	16:00	12:30		
16				출근								0.00
17				퇴근								
18				출근								0.00
19				퇴근								
20				출근								0.00
21				퇴근								
22				출근								0.00

▶ 전처리가 필요한 엑셀 자료

이 자료로 월간, 분기 단위의 집계 또는 보고서 작성 작업을 해야 한다면 난감해진다.

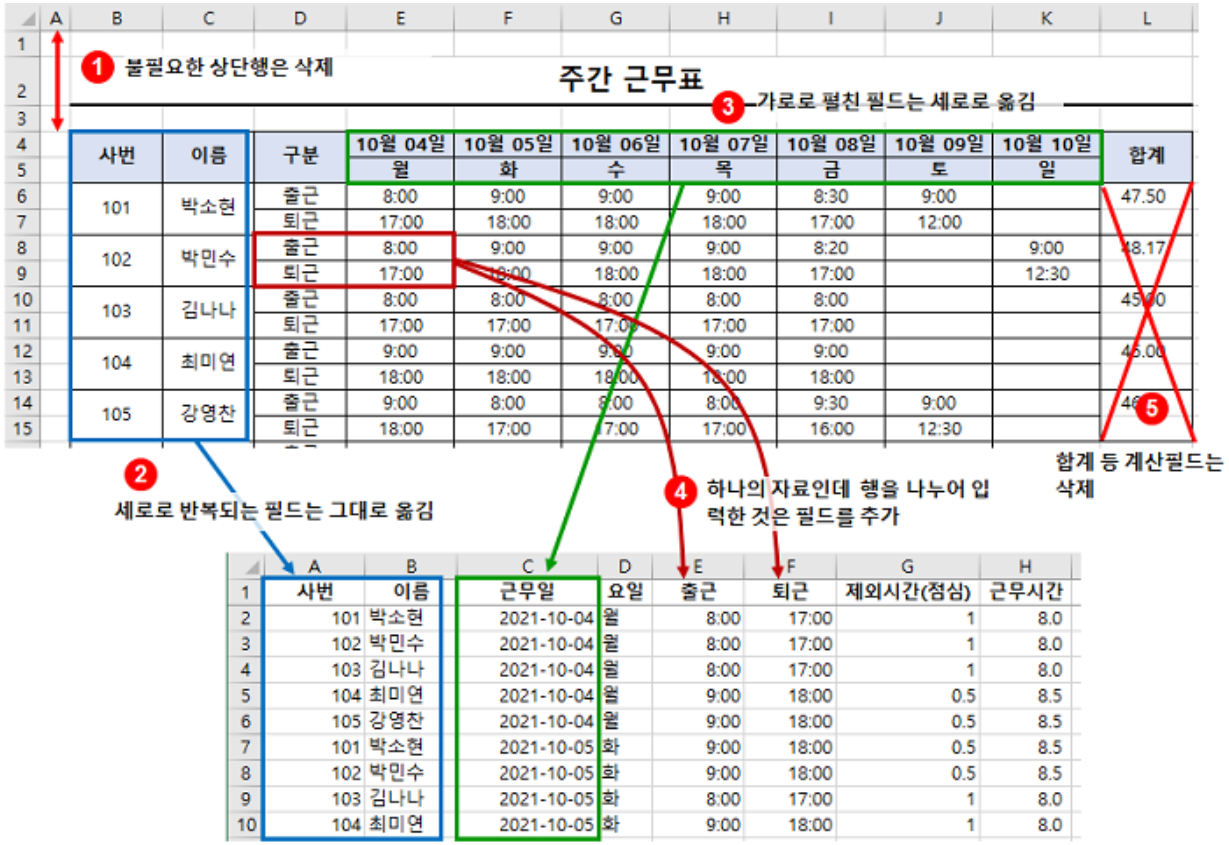
시간을 상당히 들여서 수작업하거나 복잡하고 꼬인 수식을 쓰든지 해야 한다.

- 단기간만 쓸 때는 금방 만들수 있고, 간단해서 좋겠지만 몇 달 지나 시트가 많아지면 자료에 파묻힌다.
- 주간 단위 관리를 위한 포맷이므로 월간,분기 집계/보고 등 주간 단위 이외의 자료 처리는 많은 노력을 들여야 한다.
- 주간단위 근무 시간 합계 같은 간단한 집계도 배열 수식을 써야한다({=SUM((E7:K7)-(E6:K6))*24})
- 점심시간 계산은 되지도 않으니 필드를 추가하거나 근무시간에서 일괄적으로 1을 빼는 수식을 작성해야 하고, 혹시 연장 근무 계산이 필요하면 다 뜯어 고쳐야 한다.

	A	B	C	D	E	F	G	H
1	사번	이름	근무일	요일	출근	퇴근	제외시간(점심)	근무시간
2	101	박소현	2021-10-04	월	8:00	17:00	1	8.0
3	102	박민수	2021-10-04	월	8:00	17:00	1	8.0
4	103	김나나	2021-10-04	월	8:00	17:00	1	8.0
5	104	최미연	2021-10-04	월	9:00	18:00	0.5	8.5
6	105	강영찬	2021-10-04	월	9:00	18:00	0.5	8.5
7	101	박소현	2021-10-05	화	9:00	18:00	0.5	8.5
8	102	박민수	2021-10-05	화	9:00	18:00	0.5	8.5
9	103	김나나	2021-10-05	화	8:00	17:00	1	8.0
10	104	최미연	2021-10-05	화	9:00	18:00	1	8.0
11	105	강영찬	2021-10-05	화	8:00	17:00	1	8.0
12	101	박소현	2021-10-06	수	9:00	18:00	1	8.0
13	102	박민수	2021-10-06	수	9:00	18:00	1	8.0

▶ 어떤 전처리가 필요한가?

① 상단 1~3행은 불필요하므로 삭제한다. 머리글이 1행, 1열([A1]셀)부터 시작하도록하고 제목은 시트 이름으로 대신한다.
② 세로로 반복되는 사번, 이름 필드는 그대로 세로로 옮긴다.
③ 날짜와 함께 요일이 가로로 펼쳐진 부분은 확장에 문제가 있으므로(1주일이 아니라 1개월을 관리해야 한다면 31개까지 열을 추가해야 함) 근무일 필드와 요일 필드를 추가하고 데이터는 세로로(행단위)로 추가될 수 있도록 한다.
④ 출근시간과 퇴근시간은 하나의 자료(행단위)에 기록되어야 하는데 2개의 행으로 나누어져 있다. 출근, 퇴근 필드를 추가한다.
⑤ 합계 등 계산 필드는 삭제한다. 계산 필드는 집계, 분석, 보고용 시트에서 따로 계산할 수 있으므로 불필요한 필드이다.



사번	이름	구분	10월 04일 월	10월 05일 화	10월 06일 수	10월 07일 목	10월 08일 금	10월 09일 토	10월 10일 일	합계
101	박소현	출근	8:00	9:00	9:00	9:00	8:30	9:00		47.50
		퇴근	17:00	18:00	18:00	18:00	17:00	12:00		
102	박민수	출근	8:00	9:00	9:00	9:00	8:20		9:00	48.17
		퇴근	17:00	18:00	18:00	18:00	17:00		12:30	
103	김나나	출근	8:00	8:00	8:00	8:00	8:00			45.00
		퇴근	17:00	17:00	17:00	17:00	17:00			
104	최미연	출근	9:00	9:00	9:00	9:00	9:00			46.00
		퇴근	18:00	18:00	18:00	18:00	18:00			
105	강영찬	출근	9:00	8:00	8:00	8:00	9:30	9:00		46
		퇴근	18:00	17:00	17:00	17:00	16:00	12:30		

사번	이름	근무일	요일	출근	퇴근	제외시간(점심)	근무시간
101	박소현	2021-10-04	월	8:00	17:00	1	8.0
102	박민수	2021-10-04	월	8:00	17:00	1	8.0
103	김나나	2021-10-04	월	8:00	17:00	1	8.0
104	최미연	2021-10-04	월	9:00	18:00	0.5	8.5
105	강영찬	2021-10-04	월	9:00	18:00	0.5	8.5
101	박소현	2021-10-05	화	9:00	18:00	0.5	8.5
102	박민수	2021-10-05	화	9:00	18:00	0.5	8.5
103	김나나	2021-10-05	화	8:00	17:00	1	8.0
104	최미연	2021-10-05	화	9:00	18:00	1	8.0

▶ 데이터 전처리 기본 규칙

○ **결측값(null)을 처리한다.**
데이터의 중간에 빈 셀, 빈 행이 있으면 안된다. 특히 셀 병합은 결측이 발생하므로 주의한다.

○ **한 필드의 데이터 유형을 통일한다.**
날짜 필드이라면 날짜 형식으로만 넣는다. 텍스트로 '9월 15일'의 형식으로 입력하면 안된다.

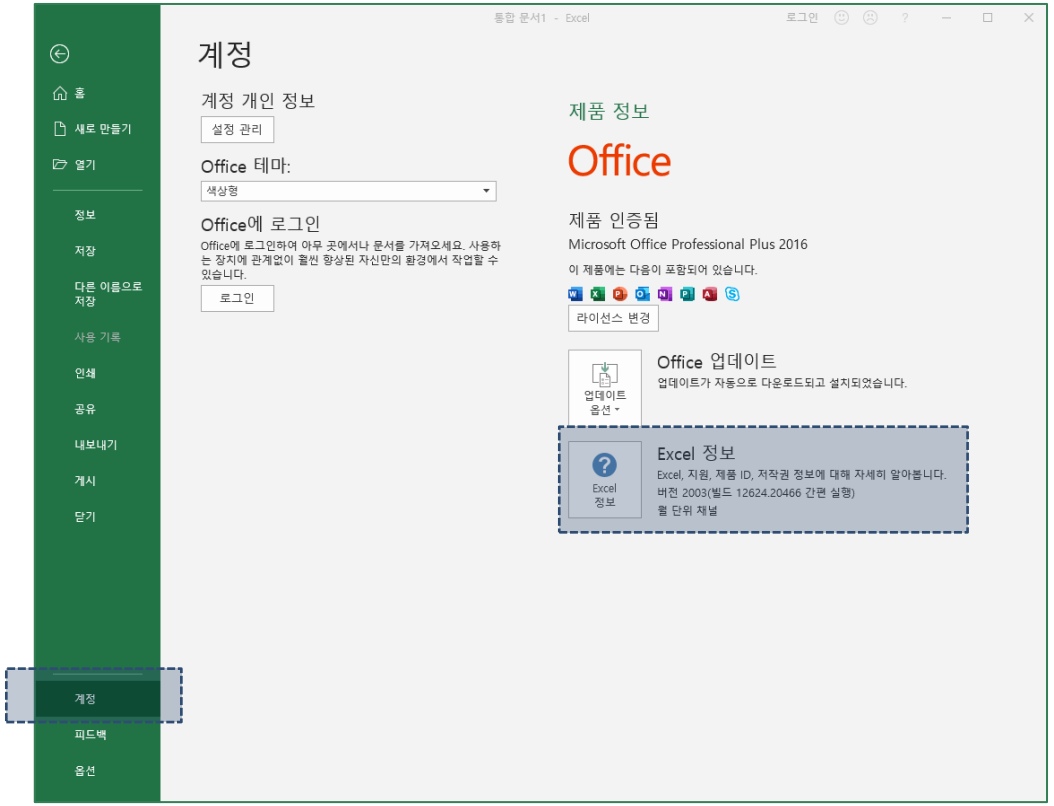
○ **한 필드에는 하나의 정보만 담는다.**
"서울 20대 여성"처럼 한 열에 여러 정보를 담으면 성별에 따라, 나잇대에 따라, 지역에 따라 데이터를 집계하고 분석하기 어렵다.

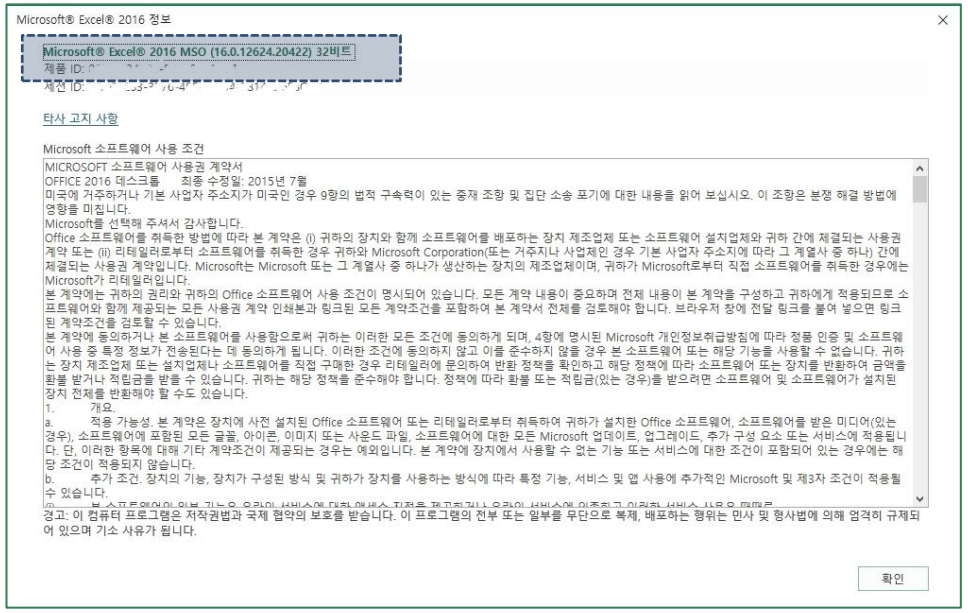
○ **한 필드의 데이터는 같은 측정 단위로, 유효한 값을, 일관성 있게 기록한다.**
학생마다 누군가의 키는 cm 단위로, 누군가는 m 단위로 적어서는 안된다.
학생의 키 데이터에 500cm 등 유효하지 않은 값이 담기지 않게 주의한다.
같은 뜻을 가진 데이터(월요일, 월, 月, MON, Monday 등)는 한 가지로 표현(월요일)한다.
쉼표 스타일(1,000 단위 구분기호), 알파벳 대소문자, 보이지 않는 앞뒤 공백 등은 모두 다른 데이터이다.

○ **값만 데이터로 취급한다.**
폰트 크기, 셀 배경색 등은 정렬 기준으로 삼을 수 있지만 데이터의 처리와 분석에는 유용하지 못하다.
수식이나 하이퍼링크 등을 제외한 값만 데이터로 취급한다.

▶ 파워쿼리 활성화

- [파일] → [계정] → [EXCEL 정보]를 클릭하면 사용하는 엑셀의 버전을 확인할 수 있다.



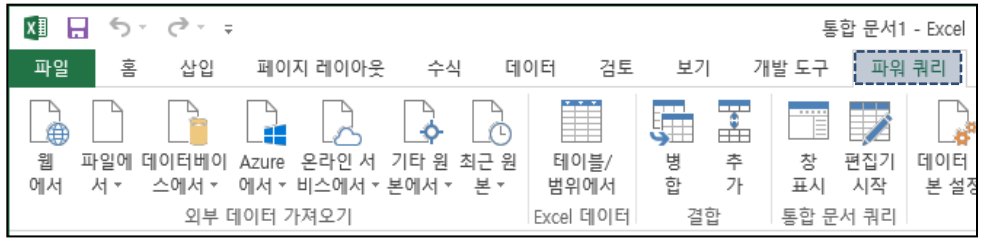


▶ 파워쿼리 활성화

- 엑셀 버전별 파워쿼리 지원
  - 오피스 2010 Professional Plus 이상, 오피스 2013 : add-on 설치 후 사용할 수 있다.
  - 오피스 2016 : 엑셀의 기본 기능으로 내장되어 있다. (업데이트- https://www.microsoft.com/ko-KR/download/details.aspx?id=56547)
  - MAC 및 student 등 일부 엑셀 버전은 파워쿼리를 온전히 지원하지 않으므로 Power BI를 이용해야 한다.

- Add-on 설치하는 경우
  - 파워쿼리를 다운로드 한다. (https://www.microsoft.com/ko-KR/download/details.aspx?id=39379)
  - 설치하면 엑셀 상단 메뉴에 [파워쿼리] 탭이 나타난다.

- 기본 기능으로 내장된 경우
  - [데이터] 탭에 파워쿼리가 포함되어 있다.



▶ 파워 쿼리의 이해

- 다양한 유형의 데이터를 검색, 연결, 결합
- 분석 가능한 형태의 데이터로 가공

테이블 형식

매출일자	성명	부서	직위	지역	온라인	오프라인	총매출액
2017-01-03	김형진	전산팀	차장	서울	8,200	9,300	17,500
2017-01-04	김소미	기획팀	과장	서울	5,400	6,190	11,590
2017-01-06	송윤희	홍보팀	차장	광주	4,290	6,190	10,480
2017-01-06	김소미	기획팀	과장	서울	5,130	5,880	11,010
2017-01-07	김찬혁	인사팀	차장	서울	7,230	6,190	13,420
2017-01-08	송윤희	홍보팀	차장	광주	6,870	5,880	12,750
2017-01-09	김찬혁	인사팀	차장	서울	7,100	5,100	12,200
2017-01-11	유가을	영업팀	과장	서울	5,800	6,960	12,760
2017-01-13	한재원	영업팀	사원	서울	5,800	6,600	12,400
2017-01-14	유가을	영업팀	과장	서울	5,510	6,610	12,120
2017-01-15	한재원	영업팀	사원	서울	5,510	7,210	12,720
2017-01-17	김덕훈	영업팀	차장	서울	7,520	6,830	14,350
2017-01-18	안정훈	전산팀	대리	울산	6,410	7,980	14,390
2017-01-19	김덕훈	영업팀	차장	서울	7,520	6,830	14,350

		서울			인천			대전		
		온라인	오프라인	기타	온라인	오프라인	기타	온라인	오프라인	기타
3	2010	67,173	85,406	46,293	19,047	57,613	21,552	96,976	80,503	48,032
4	2011	36,278	93,715	47,093	51,747	57,480	32,498	33,194	74,512	25,598
5	2012	87,006	68,573	24,783	59,901	97,475	76,355	81,936	12,854	76,708
6	2013	86,989	96,044	70,376	99,951	86,621	87,060	29,655	99,114	53,934
7	2014	43,234	88,941	37,129	62,387	51,832	63,308	56,894	20,696	95,629
8	2015	54,894	18,198	82,046	44,270	26,417	18,200	41,555	28,581	94,893
9	2016	33,394	80,252	82,305	52,182	79,627	22,871	53,535	20,019	27,480
10	2017	50,700	42,066	93,349	42,996	66,091	50,279	39,730	86,646	29,552

Category	Manufacturer	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	2014 Total
Mix	Abbas	341	442	703	699	772	697	555	518	539	521	434	521	6,742
	Aliqui	230	334	617	819	955	829	596	603	411	272	251	340	6,257
	Currus	407	665	1,307	1,511	1,608	1,317	1,039	902	680	587	471	593	11,087
	Natura	202	293	545	520	504	419	356	364	310	246	184	239	4,182
	Pirum	467	612	1,320	1,333	1,453	1,329	942	972	806	559	577	464	10,834
	Pomum	4	6	10	7	13	11	11	9	4	11	5	4	95
	Quibus	284	293	410	495	410	353	280	254	194	209	238	318	3,738
	Victoria	18	27	74	64	67	75	51	45	33	33	23	20	530
Mix Total		1,953	2,672	4,986	5,448	5,782	5,030	3,830	3,667	2,977	2,438	2,183	2,499	43,465

Category	Manufacturer	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	2014 Total
Rural	Abbas	13	26	31	42	18	10	20	25	18	21	9	33	266
	Aliqui	7,074	8,256	12,808	14,375	14,495	11,358	8,724	9,236	7,588	7,418	9,032	21,971	132,335
	Currus	1,854	2,473	3,479	3,414	3,257	2,861	2,552	2,285	2,073	2,163	2,701	6,880	35,992
	Fama	2	2				8		4		2			18
	Natura	11,414	13,558	20,787	19,659	19,107	16,633	15,439	14,565	13,855	13,985	18,065	37,557	214,624
	Pirum	4,254	4,299	6,619	6,173	5,391	4,668	3,635	4,353	3,861	3,716	4,012	7,305	58,286
	Pomum	12	12	44	26	38	24	16	34	38	26	28	26	324
	Quibus	3,304	3,104	4,056	4,191	3,531	2,912	2,017	2,621	3,041	2,825	2,854	3,472	37,928
	VanArsdel	5	13	27	8	7	2	16	7	6	4	6	5	106
Rural Total		27,932	31,743	47,851	47,888	45,844	38,476	32,419	33,130	30,480	30,160	36,707	77,249	479,879

구매 내역 확인서

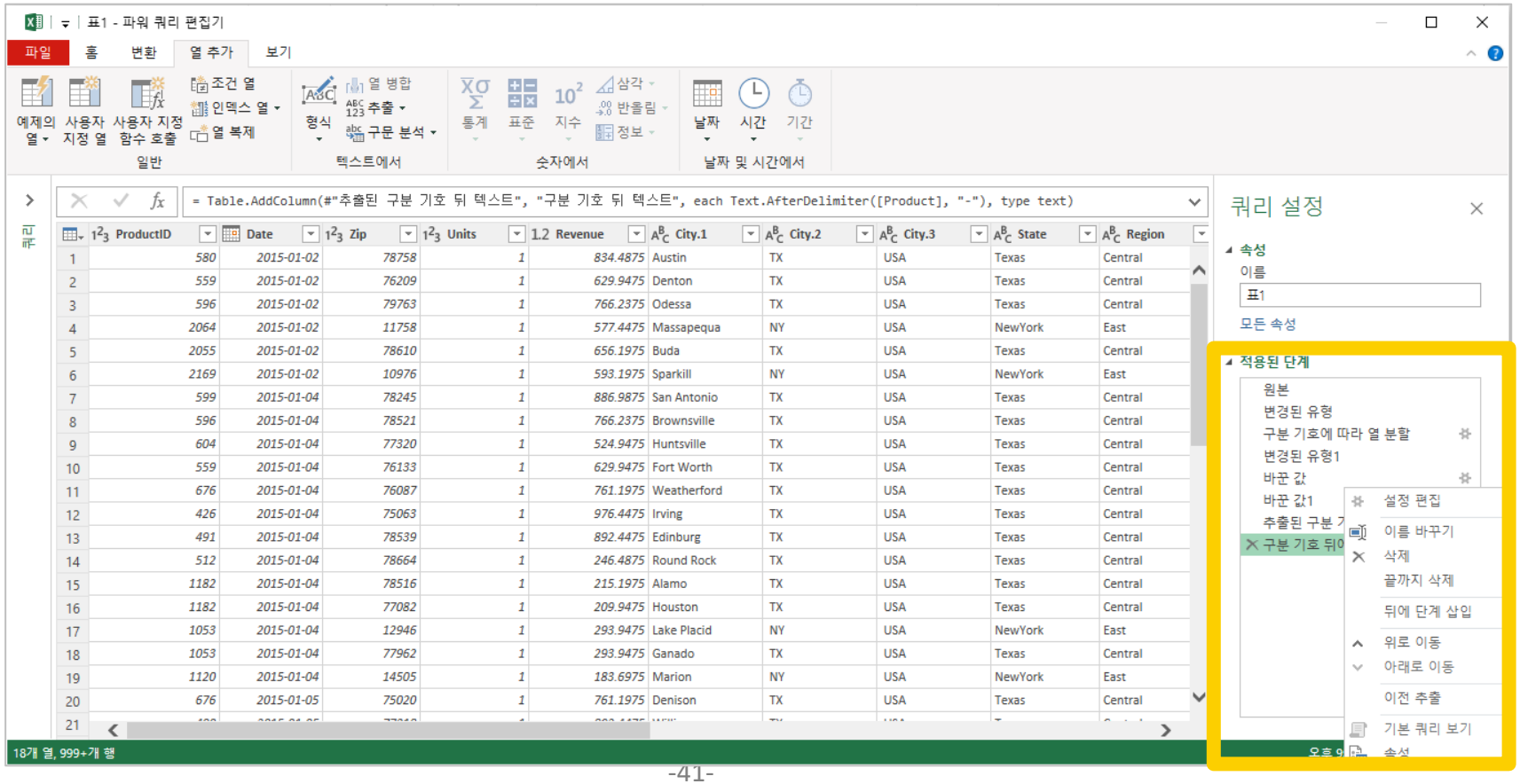
담당	팀장	이사

작성일자 : 2017-10-10 작성자 : 송윤희

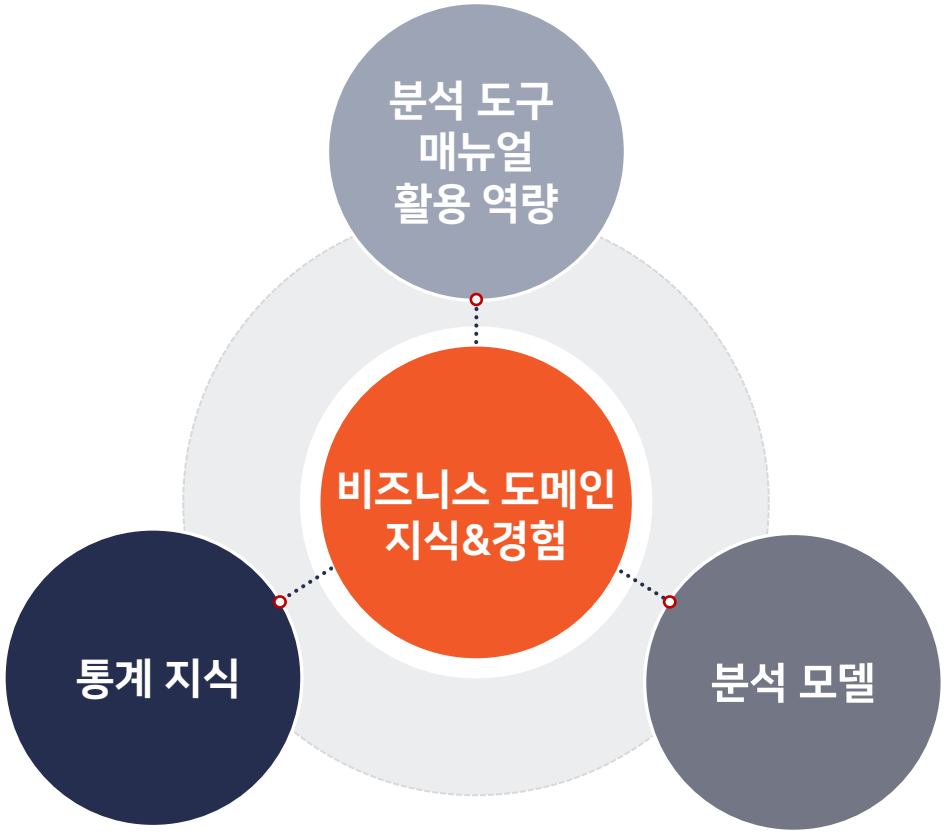
고객코드	고객명	전화번호	구매수량	고객등급	배송지역	배송료
A02	삼일	010-2222-2222	37	실버	서울	5,000
B01	신영상사	051-575-5776	52	골드	경기	5,500
B03	정금상사	041-932-3778	85	프리미엄	대전	8,000
C01	양정물산	02-444-2971	15	일반	부산	12,000
A04	태강	010-8888-8888	60	골드	광주	12,000
B02	경성트레이딩	031-776-4568	76	골드	제주도	15,000
C05	유미백화점	031-768-7688	92	프리미엄	서울	5,000

▶ 파워 쿼리 편집기

- 쿼리 편집기의 편집 내용은 자동으로 기록
- 원본 데이터 변경 후 [새로 고침] 시 기록된 편집 내용 자동으로 적용
- 저장된 작업 단계의 편집, 이동, 삭제 등 관리 가능
- [쿼리 설정] 창 - [적용된 단계]



▶ 비즈니스 데이터 분석을 위한 필수 역량



| 분석 도구 매뉴얼 활용 역량

다양한 데이터 기획, 수집, 처리, 관리, 분석, 시각화 도구를 다루는 기술

Excel, R, Python, SPSS, SAS, GA, AzureML, Tableau, SQL, Power Platform 등

| 통계 지식

데이터의 수집과 측정, 처리, 분석, 해석을 위한 통계분석 지식

피어슨 통계(기술통계, 확률론, 추론통계), 베이즈 통계(조건부 확률, 축차합리성) 등

| 비즈니스 도메인 지식 & 경험

데이터 사이언스 적용 대상 실무 분야의 축적된 고유한 경험과 지식

도소매 판매, 물류, IT, 광고 등 각 분야의 주요 지표와 의미 해석, 활용 등의 노하우

| 분석 모델

합리적인 결론 도출을 위한 데이터의 분석 방법 설계 역량

비즈니스 분석 모델(ARM, AARRR 등), 연관성 규칙, 선형계획법, 시뮬레이션,

분류와 군집화(K-means Cluster, Bagging, Random Forest, Boosting 등) 등

▶ 데이터 분석 프로세스

데이터 수집
- 적정 규모의 샘플 데이터 표집
- 다양한 취합 방식
  (crawling, survey, tagging, logging 등)

데이터 전처리
- 데이터 유형, 척도 등 정리
- 결측값 처리
- 테이블(table) 형태 변환
- 표준화와 정규화
- …

Clean Dataset

탐색적 데이터 분석
EDA, Exploratory Data Analysis
- 기술통계법
- 일변량 데이터 탐색
- 다변량 데이터 탐색
- 이상치 탐색
- …

분석 모델 & 알고리즘 적용
- 회귀 예측 모델
- 분류와 군집화
- 시계열 분석
- 지도, 비지도, 강화 학습
- 이미지, 감성 분석
- 시뮬레이션
- …

데이터 기반 커뮤니케이션
- 데이터 분석 결과의 시각화 보고서를 활용한 데이터 기반 커뮤니케이션
- …

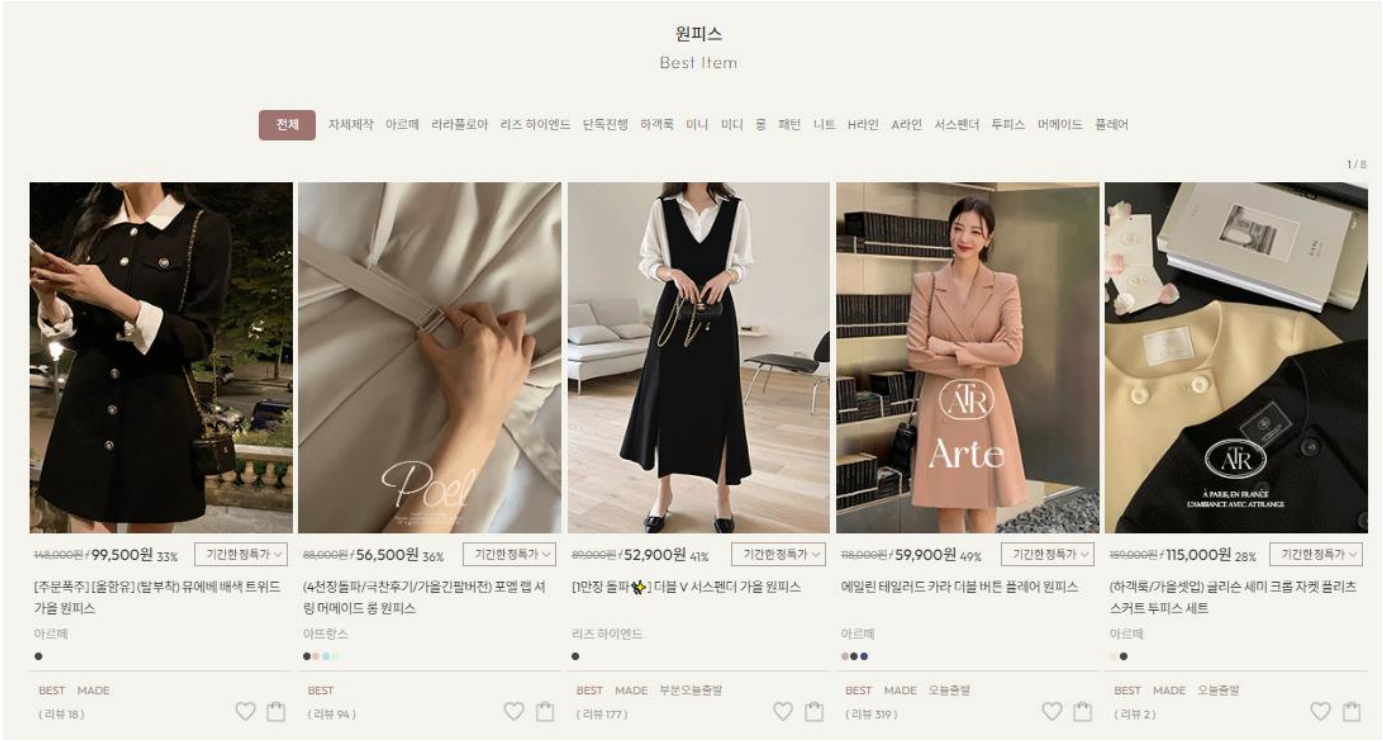
Decision Making

현실세계
Business Question

▶ 데이터 분석 프로세스

Quiz

동절기 매출이 부진한 원피스 쇼핑몰은 어떤 제품군을 팔아야 할까?



Source: attrangs.com

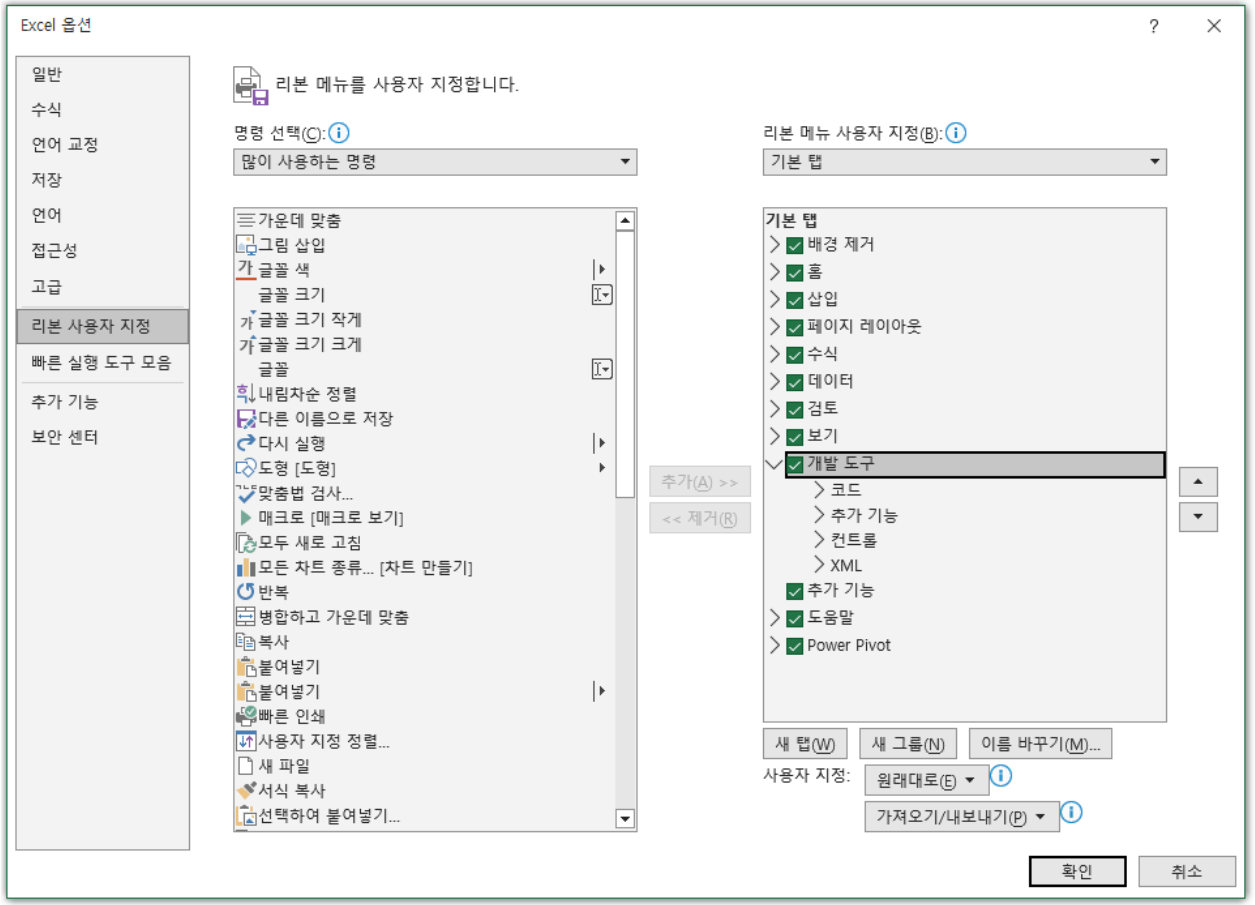
▶ 엑셀 매크로와 VBA의 이해

- 매크로는 자주 반복되는 여러 단계 작업을 하나로 묶어 실행하는 기능이다.
  하나로 묶인 작업은 엑셀에서 메뉴 버튼, 단축키, 도형이나 이미지와의 연결, 양식 컨트롤과의 연결,
  VBA 편집기로 실행 조건과 명령어 코딩 등을 통해 쉽게 반복 실행된다.
- 매크로는 Microsoft Office 에 내장된 프로그래밍 언어인 VBA(Visual Basic for Application)로 만들어진다.
  VBA와 매크로는 모든 오피스 프로그램에 포함되어 있지만, 매크로 기록 기능은 EXCEL과 Words에만 있다.
- [Alt] + [F8] 단축키로 매크로 메뉴를 실행할 수 있다.

▶ 개발 도구 활성화 하기

① [파일] → [옵션] → [리본 사용자 지정]

② [리본 메뉴 사용자 지정] → [개발 도구] 체크



THANK YOU

마소캠퍼스 www.masocampus.com
이메일 문의 biz@masocampus.com
전화 문의 02-6080-2022

Actionable Content
MASO CAMPUS