

엑셀로 이해하는 인공지능

2022. 09. 29.

허주용(huhjuang@gmail.com)



허주용
(huhjuang
@gmail.com)

▣ 근무 경력

- 현, 다비(DAVI) 대표
- 전, (주)애드웹커뮤니케이션 사업개발팀 팀장 / 빅데이터 분석, 솔루션 기획
- 전, (주)엔에이치앤커머스(구, 고도소프트) 이사 / 개발, 신사업 총괄
- 전, (주)가비아 인프라사업부 부장 / 솔루션 기획/PM, 마케팅, CRM
- 전, (주)쌍용건설 투자개발실 / SOC사업 타당성 분석 및 대관영업

▣ 강의 경력

- 메디치 : 인공지능(AI) 기반 빅데이터 분석 전문가 양성 과정 외
- 휴먼교육센터 : AI데이터플랫폼을 활용한 빅데이터 분석전문가 과정
- 멀티캠퍼스 : KDT 융복합 프로젝트 5개 과정 외
- 한국공인회계사회 BI 솔루션을 활용한 회계실습 강의
- 연세대학교 상남경영원 빅데이터 재무분석 전문가과정 강의

▣ 주요 프로젝트 및 기타사항

- 자동화 : e커머스 데이터 분석 및 리포팅 업무 처리 자동화 등
- 분석 : 볼링스코어 인식, 로그데이터 유입 분류 분석, 사주 연관성 분석, 경마 우승마 예측 분석, e커머스 매출 분석 등
- 텍스트마이닝 : 대통령 연설문 분석, 인터넷정보보호 10대 이슈, 노동정책 반응, 노동감독관 이슈, 에듀테크 이슈 등
- (주)엔에이치앤커머스 e나무 쇼핑몰 빌더 개발 총괄
- (주)가비아 솔루션 기획 및 PM : 도메인솔루션, 홈페이지빌더, 쇼핑몰빌더, 하이웍스 등
- [저서]클릭클릭! 코딩없이 시작하는 엑셀 크롤링(위키독스)
- [저서]클릭클릭! 엑셀로 이해하는 인공지능(위키독스)

CONTENTS

I 엑셀 고급함수

- 1. 상대참조, 절대참조
- 2. vlookup, iferror
- 3. hlookup
- 4. sumif, countif, sumproduct

II 엑셀로 이해하는 기초통계량과 머신러닝

- 1. EDA
- 2. 지도학습
- 3. 비지도학습
- 4. 딥러닝
- 5. 기타

I. 엑셀 고급함수

1. 상대참조, 절대참조
2. vlookup, iferror
3. hlookup
4. sumif, countif, sumproduct

1. 상대참조, 절대참조

A	B	C	D	E	F	G
1						
2	[매출]				* 수식보기	
3	지점	수수료율	가방	옷		
4	계		350	150		
5	서울	5%	0	30		
6	부산	10%	150	50		
7	인천	8%	200	70		
8						
9	[상대참조]					
10	지점	수수료율	가방	옷	가방수수료	옷수수료
11			350	150	0	0
12	서울	5%	0	30	0	0
13	부산	10%	150	50	0	0
14	인천	8%	200	70	0	0

A	B	C	D	E	F	G
9	[상대참조]					
10	지점	수수료율	가방	옷	가방수수료	옷수수료
11			=D4	=E4	=F4	=G4
12	서울	=C5	=D5	=E5	=F5	=G5
13	부산	=C6	=D6	=E6	=F6	=G6
14	인천	=C7	=D7	=E7	=F7	=G7

1. 상대참조, 절대참조

	A	B	C	D	E	F	G
16	[절대참조]						
17	지점	수수료율	가방	옷	가방수수료	옷수수료	
18			5%	5%	5%	5%	
19	서울	5%	5%	5%	5%	5%	
20	부산	5%	5%	5%	5%	5%	
21	인천	5%	5%	5%	5%	5%	

	A	B	C	D	E	F	G
16	[절대참조]						
17	지점	수수료율	가방	옷	가방수수료	옷수수료	
18			=\$C\$5	=\$C\$5	=\$C\$5	=\$C\$5	
19	서울	=\$C\$5	=\$C\$5	=\$C\$5	=\$C\$5	=\$C\$5	
20	부산	=\$C\$5	=\$C\$5	=\$C\$5	=\$C\$5	=\$C\$5	
21	인천	=\$C\$5	=\$C\$5	=\$C\$5	=\$C\$5	=\$C\$5	

1. 상대참조, 절대참조

A	B	C	D	E	F	G	H	I	J	K	L	M	N
23	[열고정]							[열고정 설정]					
24	지점	수수료율	가방	옷	가방수수료	옷수수료		지점	수수료율	가방	옷	가방수수료	옷수수료
25			350	150	-					350	150	31	12
26	서울	5%	-	30	-	-		서울	5%	-	30	-	2
27	부산	10%	150	50	15	7,500		부산	10%	150	50	15	5
28	인천	8%	200	70	16	14,000		인천	8%	200	70	16	6

A	B	C	D	E	F	G	H	I	J	K	L	M	N
23	[열고정]						[열고정 설정]						
24	지점	수수료율	가방	옷	가방수수료	옷수수료		지점	수수료율	가방	옷	가방수수료	옷수수료
25		=SUM(D26:D28)	=SUM(E26:E28)	=F4					=SUM(K26:K28)	=SUM(L26:L28)	=SUM(M26:M28)	=SUM(N26:N28)	
26	서울	=C5	=D5	=E5	=D26*C26	=E26*D26		서울	=C5	=D5	=E5	=K26*\$J26	=L26*\$J26
27	부산	=C6	=D6	=E6	=D27*C27	=E27*D27		부산	=C6	=D6	=E6	=K27*\$J27	=L27*\$J27
28	인천	=C7	=D7	=E7	=D28*C28	=E28*D28		인천	=C7	=D7	=E7	=K28*\$J28	=L28*\$J28

1. 상대참조, 절대참조

A	B	C	D	E	F	G	H	I	J	K	L
30	[행고정 : 비율구하기]							[행고정 : 비율구하기 설정]			
31	지점	수수료율	가방	옷				지점	수수료율	가방	옷
32			350	150						350	150
33	서울	5%	0%	20%				서울	5%	0%	20%
34	부산	10%	#DIV/0!	167%				부산	10%	43%	33%
35	인천	8%	133%	140%				인천	8%	57%	47%

A	B	C	D	E	F	G	H	I	J	K	L
30	[행고정 :							[행고정 : 비율			
31	지점	수수료율	가방	옷				지점	수수료율	가방	옷
32			=D4	=E4					=D4	=E4	
33	서울	=C12	=D5/D4	=E5/E4				서울	=C12	=D5/D\$4	=E5/E\$4
34	부산	=C13	=D6/D5	=E6/E5				부산	=C13	=D6/D\$4	=E6/E\$4
35	인천	=C14	=D7/D6	=E7/E6				인천	=C14	=D7/D\$4	=E7/E\$4

2. vlookup, iferror

	A	B	C	D	E	F	G
1		- vlookup으로 다른 테이블의 값을 가져오기					
2		- if로 수식 오류 해결하기					
3		- iferror로 수식 오류 해결하기					
4							
5	[매출]			if	iferror	iferror	
6	지점	매출	비용	비용/매출	비용/매출	비용/매출	
7	서울	0	30	ERROR	ERROR		
8	부산	150	50	33%	33%	33%	
9	인천	200	70	35%	35%	35%	
10							
11	[비용]						
12	지점	비용					
13	부산	50					
14	인천	70					
15	서울	30					

	A	B	C	D	E	F	G
5	[매출]				if		iferror
6	지점	매출	비용		비용/매출	비용/매출	비용/매출
7	서울	0	=VLOOKUP(B7,\$B\$13:\$C\$15,2,0)	=IF(ISERROR(D7/C7), "ERROR", D7/C7)	=IFERROR(D7/C7,"ERROR")	=IFERROR(D7/C7,"")	
8	부산	150	=VLOOKUP(B8,\$B\$13:\$C\$15,2,0)	=IF(ISERROR(D8/C8), "ERROR", D8/C8)	=IFERROR(D8/C8,"ERROR")	=IFERROR(D8/C8,"")	
9	인천	200	=VLOOKUP(B9,\$B\$13:\$C\$15,2,0)	=IF(ISERROR(D9/C9), "ERROR", D9/C9)	=IFERROR(D9/C9,"ERROR")	=IFERROR(D9/C9,"")	

3. hlookup

	A	B	C	D	E
1	- hlookup으로 다른테이블의 값을 가져오기				
2					
3	[매출]				
4	지점	서울	부산	인천	
5	매출		0	150	200
6	비용		30	50	70
7					
8					
9	[비용]				
10	지점	부산	인천	서울	
11	비용		50	70	30

	A	B	C	D	E
3	[매출]				
4	지점	서울		부산	
5	매출	0		150	인천
6	비용	=HLOOKUP(C4,\$C\$10:\$E\$11,2,FALSE)		=HLOOKUP(D4,\$C\$10:\$E\$11,2,FALSE)	=HLOOKUP(E4,\$C\$10:\$E\$11,2,FALSE)

4. sumif, countif, sumproduct

A	B	C	D	E	F	G	H	I
1	- sumif로 조건을 만족하는 값 계산하기							
2	- sumproduct로 다중 조건을 만족하는 값 계산하기							
3								
4	* 매출 150이상 달성한 곳의 합계는?(sumif)			520				
5	* 매출 150이상 달성한 곳의 지역수는?(countif)			3				
6								
7	[매출]			vlookup	sumproduct	[다른방법]	vlookup	
8	월	지점	매출	비용	비용		월과지점	비용
9	1월	서울	0	30	30		1월서울	30
10	1월	부산	150	50	50		1월부산	50
11	1월	인천	200	70	70		1월인천	70
12	2월	서울	120	30	40		2월서울	40
13	2월	부산	130	50	35		2월부산	35
14	2월	인천	110	70	28		2월인천	28
15	3월	서울	170	30	40		3월서울	40
16	3월	부산	120	50	30		3월부산	30
17	3월	인천	0	70	10		3월인천	10

A	B	C	D	E	F	G
19	[비용]					
20	월	지점	월과지점	비용		
21	1월	서울	1월서울	30		
22	1월	부산	1월부산	50		
23	1월	인천	1월인천	70		
24	2월	서울	2월서울	40		
25	2월	부산	2월부산	35		
26	2월	인천	2월인천	28		
27	3월	서울	3월서울	40		
28	3월	부산	3월부산	30		
29	3월	인천	3월인천	10		
30						
31	* 매출 150이상 달성한 곳의 비용합계는?(sumproduct)			160		

4. sumif, countif, sumproduct

A	B	C	D	E	F	G
4	* 매출 150이상 달성한 곳의 합계는?(sumif)				=SUMIF(\$D\$9:\$D\$17,>=150")	
5	* 매출 150이상 달성한 곳의 지역수는?(countif)				=COUNTIF(\$D\$9:\$D\$17,>=150")	

A	B	C	D	E	F	G	H	I
7	[매출]			vlookup	sumproduct		[다른방법]	vlookup
8	월	지점	매출	비용	비용		월과지점	비용
9	1월	서울	0	=VLOOKUP(C9,\$C\$21:\$E\$29,3,FALSE)	=SUMPRODUCT((\$B\$21:\$B\$29=B9)*(\$C\$21:\$C\$29=C9),\$E\$21:\$E\$29)		=B9&C9	=VLOOKUP(H9,\$D\$21:\$E\$29,2,FALSE)
10	1월	부산	150	=VLOOKUP(C10,\$C\$21:\$E\$29,3,FALSE)	=SUMPRODUCT((\$B\$21:\$B\$29=B10)*(\$C\$21:\$C\$29=C10),\$E\$21:\$E\$29)		=B10&C10	=VLOOKUP(H10,\$D\$21:\$E\$29,2,FALSE)
11	1월	인천	200	=VLOOKUP(C11,\$C\$21:\$E\$29,3,FALSE)	=SUMPRODUCT((\$B\$21:\$B\$29=B11)*(\$C\$21:\$C\$29=C11),\$E\$21:\$E\$29)		=B11&C11	=VLOOKUP(H11,\$D\$21:\$E\$29,2,FALSE)
12	2월	서울	120	=VLOOKUP(C12,\$C\$21:\$E\$29,3,FALSE)	=SUMPRODUCT((\$B\$21:\$B\$29=B12)*(\$C\$21:\$C\$29=C12),\$E\$21:\$E\$29)		=B12&C12	=VLOOKUP(H12,\$D\$21:\$E\$29,2,FALSE)

A	B	C	D	E
19	[비용]			
20	월	지점	월과지점	비용
21	1월	서울	=B21&C21	30
22	1월	부산	=B22&C22	50
23	1월	인천	=B23&C23	70
24	2월	서울	=B24&C24	40

A	B	C	D	E	F	G
31	* 매출 150이상 달성한 곳의 비용합계는?(sumproduct)				=SUMPRODUCT((\$D\$9:\$D\$17>=150)*1,(\$F\$9:\$F\$17))	

II. 엑셀로 이해하는 기초통계량과 머신러닝

1. EDA

2. 비지도학습

3. 지도학습

4. 딥러닝

II. 엑셀로 이해하는 기초통계량과 머신러닝

14

[교육 목표]

- EDA실습을 통해 기초통계량을 이해한다
- 지도학습과 비지도학습을 구분하고 실습을 통해 그 차이를 구분한다.
- 머신러닝의 기본 개념을 엑셀 계산을 통해 이해한다

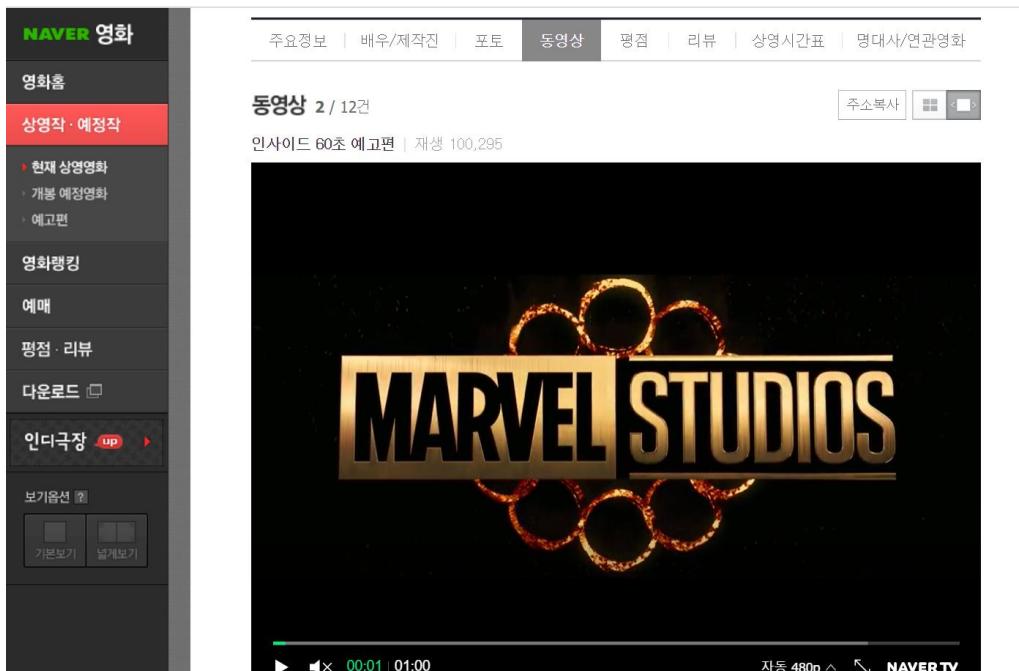
[습득 역량]

- 엑셀로 기초통계량을 구하고 시각화할 수 있다.
- 엑셀로 회귀분석, KNN, 결정트리의 원리를 이해하고 계산할 수 있다.
- 엑셀로 상관계수를 구하고 Kmeans를 통해 유사 군집을 계산할 수 있다.
- 엑셀로 퍼셉트론 모델을 계산할 수 있다.

1. EDA

15

- EDA : 데이터분석의 예고편, Preview



- 데이터 분석 분야에 대한 도메인 지식을 강화해줍니다.
- 데이터셋에 있는 오류를 알 수 있습니다.
- 전처리해야 할 부분 또는 변수선택에 대한 판단을 할 수 있게 해줍니다.
- 귀찮아서 안할 때가 많으니 자꾸 강조합니다.

* 출처 : <https://movie.naver.com/movie/bi/mi/mediaView.naver?code=187348&mid=50324#tab>

1. EDA

16

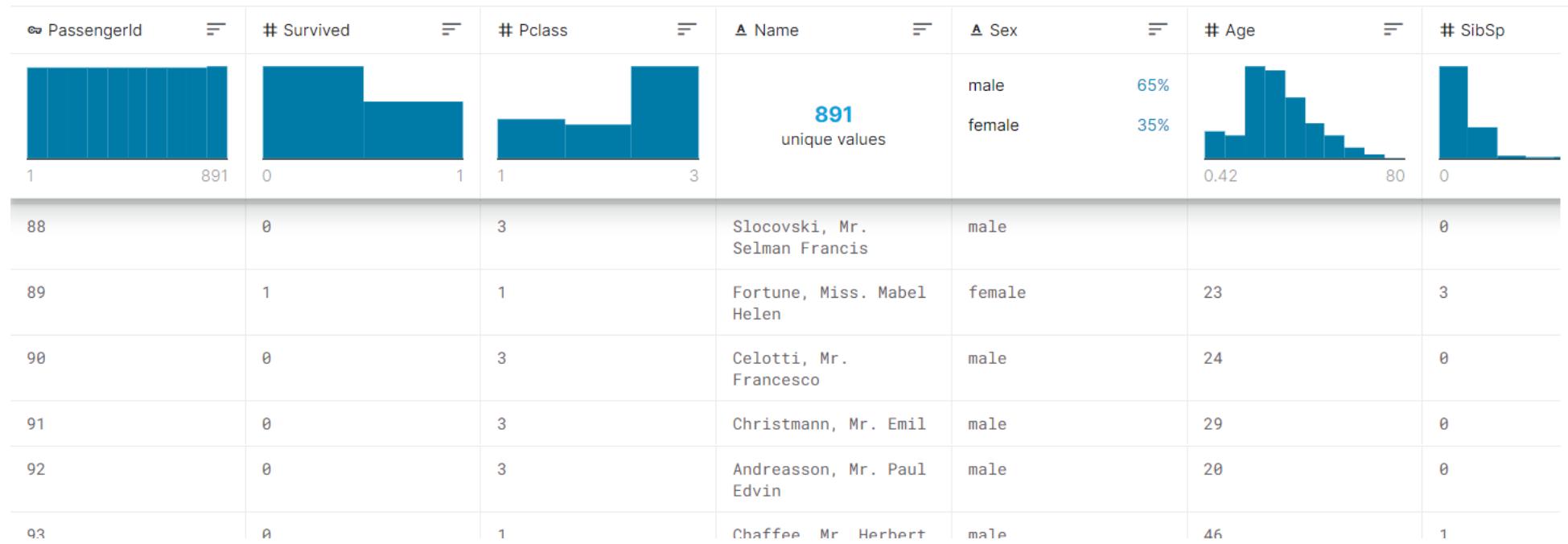
- EDA의 종류

	숫자	시각화
변수 1개	평균, 중앙값, 분산, 표준편차, 4분위, 이상치, 왜도, 첨도...	히스토그램, 박스플롯...
변수 N개	상관관계, 공분산...	산점도, 히트맵...

1. EDA

17

- 타이타닉호 생존자 데이터 사례

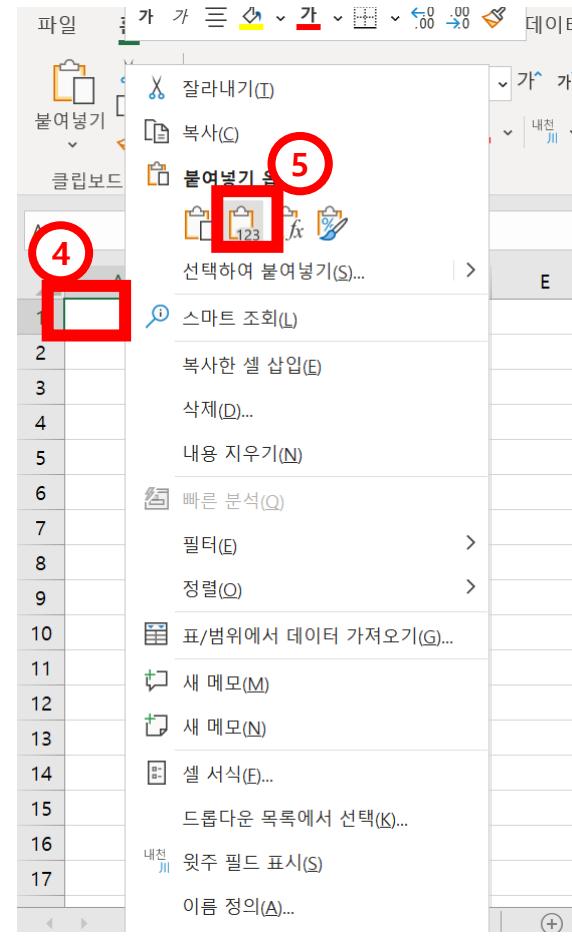


출처 : <https://www.kaggle.com/c/titanic/data?select=train.csv>

1. EDA

- 분석에 사용할 데이터를 준비합니다.
- '실거래가' sheet 이름을 '실거래가2021'로 수정하고,
- 전체를 선택한 후 복사 > 새로운 시트 만들기 > 이름을 '복사본'으로 입력 > A1 셀을 오른쪽마우스로 클릭 > 값만 붙여넣기 선택

거래금액	건축년도	년	법정동	아
80000	2002	2021	신교동	신
209000	2008	2021	사직동	광
160000	2008	2021	사직동	광
96000	2008	2021	견지동	대
32000	2003	2021	익선동	현
19700	2014	2021	연건동	이
20000	2014	2021	연건동	이
120000	1995	2021	명륜2가	아
84500	2006	2021	창신동	브
67700	1993	2021	창신동	창
127500	2004	2021	송인동	롯
11600	2012	2021	송인동	삼
13000	2013	2021	송인동	종
12500	2013	2021	송인동	종
12650	2013	2021	송인동	종
197000	2017	2021	평동	경



1. EDA

- EDA에 사용할 거래금액, 건축년도, 전용면적, 층을 제외한 나머지 년, 법정동, 아파트, 월, 일, 지번, 지역코드, 해제사유발생일, 해제여부 열을 삭제합니다.
- 열C를 선택한 후 컨트롤키를 눌러 삭제할 열을 선택한 후 오른쪽마우스를 눌러 열을 삭제합니다.

The screenshot shows a Microsoft Excel spreadsheet with 17 rows of data. The columns are labeled A through L. Row 1 contains the column headers. Rows 2 through 17 contain transaction details. Column C is highlighted with a red border and circled with number 1. A context menu is open over column C, with the '삭제(D)' option highlighted with a red box and circled with number 2. Other options in the menu include '복사(C)', '붙여넣기 옵션:', '선택하여 붙여넣기(S)...', '삽입(I)', '내용 지우기(N)', '셀 서식(E)...', '열 너비(W)...', '숨기기(H)', and '숨기기 취소(U)'. The bottom of the screen shows the ribbon tabs: 년월, 실거래가2011, 복사본, 차트, and a plus sign.

A	B	C	D	E	F	G	H	I	J	K	L	X
1	거래금액	건축년도	년	법정동	아파트	월	일	전용면적	시면	시역코드	층	해제사유별
2	80000	2002	2021	신교동	신현(101동)	8	16	84.82	6-13	11110	1	[Table]
3	209000	2008	2021	사직동	광화문풍물	8	5	163.33	9-1	11110	13	[Table]
4	160000	2008	2021	사직동	광화문풍물	8	10	158.99	9	11110	4	[Table]
5	96000	2008	2021	견지동	대성스카이	8	4	116.03	110	11110	5	[Table]
6	32000	2003	2021	익선동	현대뜨레버	8	14	48.54	55	11110	7	[Table]
7	19700	2014	2021	연건동	이화에수풀	8	2	16.98	195-10	11110	4	[Table]
8	20000	2014	2021	연건동	이화에수풀	8	29	16.98	195-10	11110	12	[Table]
9	120000	1995	2021	명륜2가	아남1	8	2	84.9	4	11110	7	[Table]
10	84500	2006	2021	창신동	브라운스톤	8	10	84.67	23-76	11110	12	[Table]
11	67700	1993	2021	창신동	창신쌍용	8	12	54.7	703	11110	3	[Table]
12	127500	2004	2021	송인동	롯데캐슬천	8	3	111.73	76	11110	17	[Table]
13	11600	2012	2021	송인동	삼전솔하임	8	6	15	296-19	11110	7	[Table]
14	13000	2013	2021	송인동	종로중흥S	8	6	17.811	202-3	11110	5	[Table]
15	12500	2013	2021	송인동	종로중흥S	8	7	17.811	202-3	11110	6	[Table]
16	12650	2013	2021	송인동	종로중흥S	8	16	17.811	202-3	11110	7	[Table]
17	197000	2017	2021	평동	경희궁자	8	2	84.614	233	11110	5	[Table]

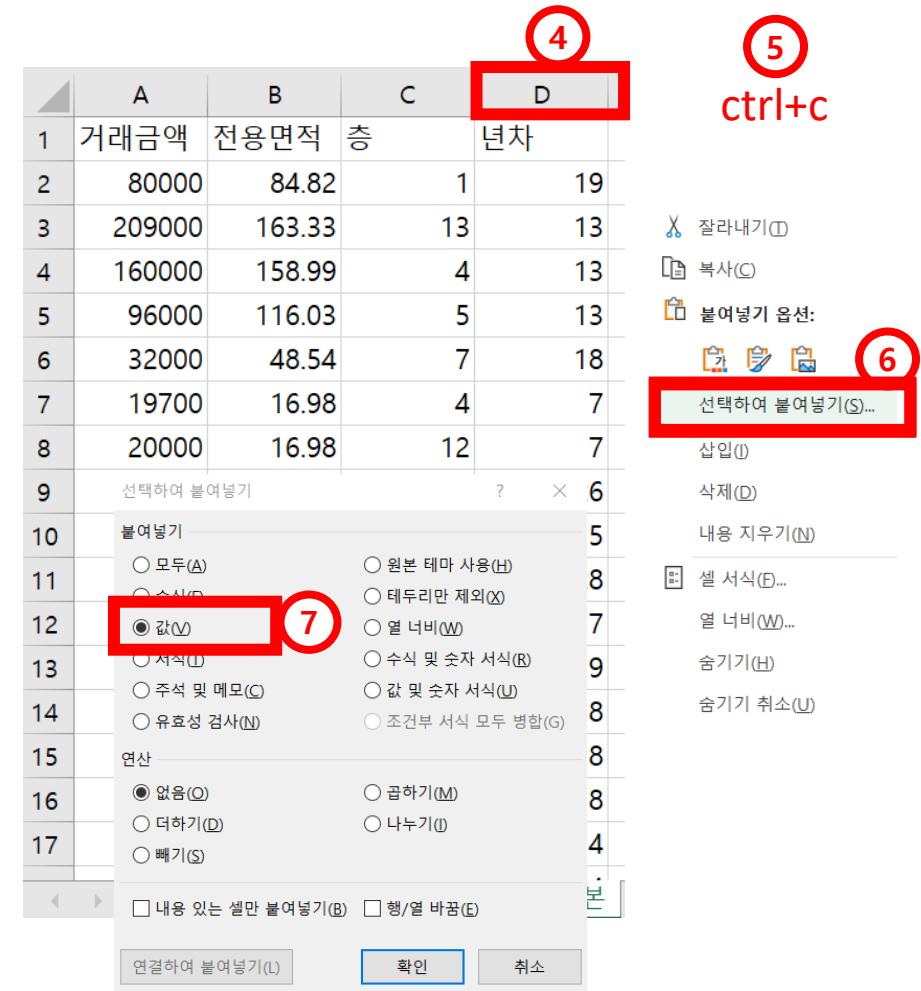
1. EDA

20

- 년차 열을 추가하여 2021년 기준으로 지은지 몇년차인지 계산합니다.
- 계산 후 년차 열을 선택 > 복사 > 오른쪽 마우스 > 값으로 붙여넣기한 후
- 건축년도 열을 삭제합니다.

The screenshot shows a Microsoft Excel spreadsheet with data from row 1 to 17. The columns are labeled A through E. Column A contains transaction amounts, column B contains construction years, column C contains conversion rates, column D contains years, and column E contains the calculated year difference. A blue arrow points from the formula in cell E2 (=2021-B2) down to cell E15. Red boxes highlight the range B2:B17 (labeled 8), the formula cell E2 (labeled 1), and the value cell E15 (labeled 3). A red arrow points from cell E15 to the context menu. The status bar at the bottom shows tabs for '연월' (Year/Month), '실거래가2011' (Actual Transaction 2011), '차트' (Chart), '복사본' (Copy), and '+'. The formula bar shows '=2021-B2'.

A	B	C	D	E
1 거래금액	건축년도	전용면적	층	년차
2 80000	2002	84.82		=2021-B2
3 209000	2008	163.33		13
4 160000	2008	158.99		13
5 96000	2008	116.03		13
6 32000	2003	48.54		18
7 19700	2014	16.98		7
8 20000	2014	16.98		7
9 120000	1995	84.9		26
10 84500	2006	84.67		15
11 67700	1993	54.7		28
12 127500	2004	111.73		17
13 116000	2012	15		9
14 130000	2013	17.811		8
15 125000	2013	17.811		8
16 126500	2013	17.811		8
17 197000	2017	84.614		4



1. EDA : 평균 외 기초통계량

- 거래금액, 전용면적, 층, 년차에 대한 평균, 중위값, 분산, 표준편차, 최소값, Q1, Q2, Q3, 최대값, IQR, 이상치(하한), 이상치(상한)을 구해봅니다.

	A	B	C	D	E	F	G	H	I	J	K
1	거래금액	전용면적	층	년차	구분	함수	거래금액	전용면적	층	년차	
2	80000	84.82	1	19	평균	average	97,886	81	8		17
3	209000	163.33	13	13	중앙값	median	87,250	84	7		17
4	160000	158.99	4	13	분산	var.s	3,466,937,518	1,749	23		91
5	96000	116.03	5	13	표준편차	stdev.s	58,881	42	5		10
6	32000	48.54	7	18	최소값	min	7,500	12	-	1	1
7	19700	16.98	4	7	Q1	quartile.inc	54,000	59	4		12
8	20000	16.98	12	7	Q2	quartile.inc	87,250	84	7		17
9	120000	84.9	7	26	Q3	quartile.inc	139,800	93	11		22
10	84500	84.67	12	15	최대값	max	375,000	239	25		50
11	67700	54.7	3	28	IQR	(Q3-Q1)	85,800	50	11		15
12	127500	111.73	17	17	이상치(하한)	Q1-1.5*IQR	-	74,700	9	-	3
13	11600	15	7	9	이상치(상한)	Q3+1.5*IQR	268,500	143	22		37
14	13000	17.811	5	8	수염(하한)		7,500	12	-	1	1
15	12500	17.811	6	8	수염(상한)		268,500	143	22		37
16	12650	17.811	7	8	왜도	skew	0.73	0.96	0.77		0.86
17	197000	84.614	5	4	첨도	kurt	1.42	1.46	0.03		1.76

1. EDA : 평균 외 기초통계량

	F	G	H	I	J	K
1	구분	함수	거래금액	전용면적	층	년차
2	평균	average	=AVERAGE(A\$2:A\$319)	=AVERAGE(B\$2:B\$319)	=AVERAGE(C\$2:C\$319)	=AVERAGE(D\$2:D\$319)
3	중앙값	median	=MEDIAN(A\$2:A\$319)	=MEDIAN(B\$2:B\$319)	=MEDIAN(C\$2:C\$319)	=MEDIAN(D\$2:D\$319)
4	분산	var.s	=VAR.S(A\$2:A\$319)	=VAR.S(B\$2:B\$319)	=VAR.S(C\$2:C\$319)	=VAR.S(D\$2:D\$319)
5	표준편차	stdev.s	=STDEV.S(A\$2:A\$319)	=STDEV.S(B\$2:B\$319)	=STDEV.S(C\$2:C\$319)	=STDEV.S(D\$2:D\$319)
6	최소값	min	=MIN(A\$2:A\$319)	=MIN(B\$2:B\$319)	=MIN(C\$2:C\$319)	=MIN(D\$2:D\$319)
7	Q1	quartile.inc	=QUARTILE.INC(A\$2:A\$319,1)	=QUARTILE.INC(B\$2:B\$319,1)	=QUARTILE.INC(C\$2:C\$319,1)	=QUARTILE.INC(D\$2:D\$319,1)
8	Q2	quartile.inc	=QUARTILE.INC(A\$2:A\$319,2)	=QUARTILE.INC(B\$2:B\$319,2)	=QUARTILE.INC(C\$2:C\$319,2)	=QUARTILE.INC(D\$2:D\$319,2)
9	Q3	quartile.inc	=QUARTILE.INC(A\$2:A\$319,3)	=QUARTILE.INC(B\$2:B\$319,3)	=QUARTILE.INC(C\$2:C\$319,3)	=QUARTILE.INC(D\$2:D\$319,3)
10	최대값	max	=MAX(A\$2:A\$319)	=MAX(B\$2:B\$319)	=MAX(C\$2:C\$319)	=MAX(D\$2:D\$319)
11	IQR	(Q3-Q1)	=H9-H7	=I9-I7	=J9-J7	=K9-K7
12	이상치(하한)	Q1-1.5*IQR	=H7-1.5*H11	=I7-1.5*I11	=J7-1.5*J11	=K7-1.5*K11
13	이상치(상한)	Q3+1.5*IQR	=H9+1.5*H11	=I9+1.5*I11	=J9+1.5*J11	=K9+1.5*K11
14	수염(하한)		=IF(H6 < H12, H12, H6)	=IF(I6 < I12, I12, I6)	=IF(J6 < J12, J12, J6)	=IF(K6 < K12, K12, K6)
15	수염(상한)		=IF(H10 > H13, H13, H10)	=IF(I10 > I13, I13, I10)	=IF(J10 > J13, J13, J10)	=IF(K10 > K13, K13, K10)
16	왜도	skew	=SKEW(A\$2:A\$319)	=SKEW(B\$2:B\$319)	=SKEW(C\$2:C\$319)	=SKEW(D\$2:D\$319)
17	첨도	kurt	=KURT(A\$2:A\$319)	=KURT(B\$2:B\$319)	=KURT(C\$2:C\$319)	=KURT(D\$2:D\$319)
18	n		=COUNT(A\$2:A\$319)			

1. EDA : 히스토그램

23

- 히스토그램은 각 변수 1개당 1개씩 그리게 됩니다.
- 해당 변수의 열머리글을 포함한 전체 데이터를 선택한 후(열머리글 선택 > CTRL+SHFT+↓)
- 삽입 > 차트에 있는 히스토그램 선택

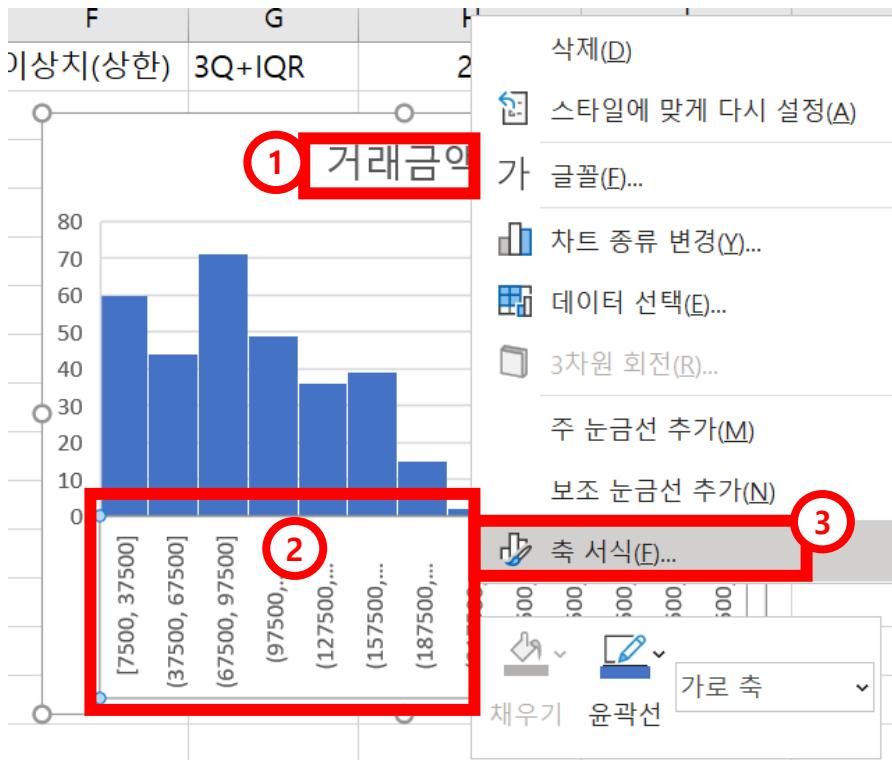
The screenshot shows the Microsoft Excel ribbon at the top. The 'Data' tab is highlighted with a red circle labeled '2'. In the 'Chart' section of the ribbon, the 'Histogram' icon is highlighted with a red circle labeled '3'. Below the ribbon, a data table is displayed in a worksheet named '거래금액'. The first column contains numerical values. A red box labeled '1' highlights the first column. The last cell in the first column, containing the value 197000, is also highlighted with a yellow background.

	A	B	C	D	E	F	G	H
1	거래금액	전용면적	층	년차	구분	함수	거래금액	
2	80000	84.82	1	19	평균	average	91	
4	209000	163.33	13	13	중앙값	median	81	
5	160000	158.99	4	13	분산	var.s	3,466,931	
6	96000	116.03	5	13	표준편차	stdev.s	58	
7	32000	48.54	7	18	최소값	min	1	
8	19700	16.98	4	7	1Q	quartile.inc	54	
9	20000	16.98	12	7	2Q	quartile.inc	81	
10	120000	84.9	7	26	3Q	quartile.inc	139	
11	84500	84.67	12	15	최대값	max	375	
12	67700	54.7	3	28	IQR	(3Q-1Q)	125	
13	127500	111.73	17	17	이상치(하한)	1Q-IQR	-	74
14	11600	15	7	9	이상치(상한)	3Q+IQR		268
15	13000	17.811	5	8				
16	12500	17.811	6	8				
17	12650	17.811	7	8				
	197000	84.614	5	4				

1. EDA : 히스토그램

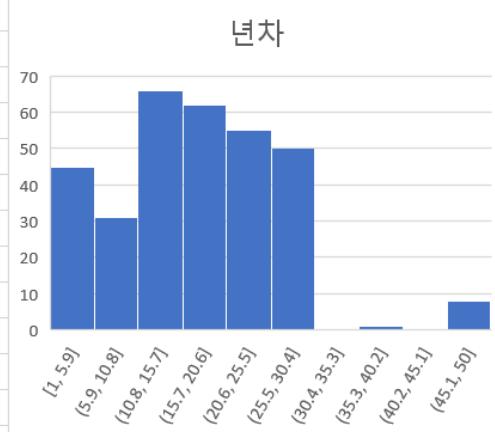
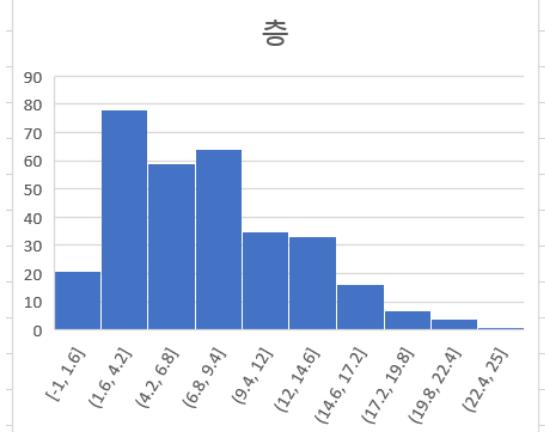
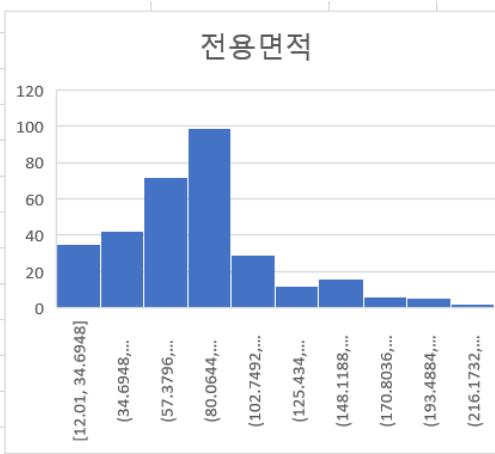
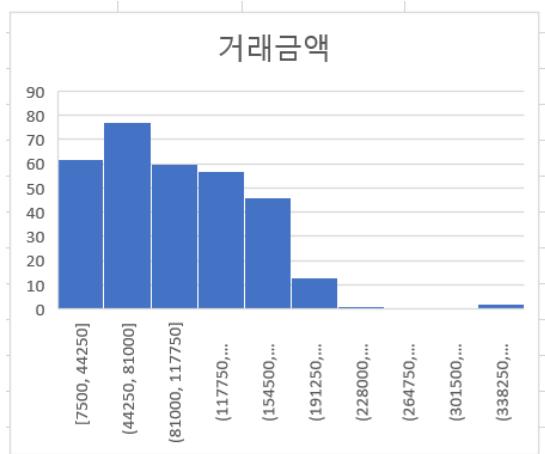
24

- 제목을 거래금액으로 수정한 후
- 히스토그램 구간수를 조정하기 위해 아래 X축 영역을 오른쪽 클릭합니다.
- 축서식을 선택한 후 축 옵션에서 계급구간 수를 10으로 수정합니다.
- 같은 방법으로 나머지 열인 전용면적, 층, 년차도 히스토그램을 만듭니다.



1. EDA : 히스토그램

- 거래금액과 층은 대체적으로 왼쪽에 몰려있고,
- 전용면적과 년자는 정규분포와 유사한 형태를 보이지만, 마찬가지로 왼쪽에 치우친 형태입니다.
- 금액이나 면적, 층, 년자 모두 최소값이 0인 경우가 많기 때문에 0에 가까울수록 몰리는 현상이 나타납니다.



1. EDA : 왜도

26

이렇게 히스토그램을 통해서 시각화하면 데이터의 좌우 쓸림을 눈으로 확인할 수 있습니다. 데이터의 좌우 쓸림을 수치로도 표현할 수 있습니다. 이 수치가 왜도(또는 비대칭도)입니다.

데이터가 좌우로 쓸려있는지 아닌지를 왜 살펴보는 것일까요? 데이터가 좌우로 쓸려있으면 분석 결과의 정확도가 떨어지거나 오류가 발생할 수 있기 때문입니다. 예를 들어 볼까요? 어떤 쇼핑몰에서 고객의 매출액 현황을 분석할 경우, 대부분 매출액이 적은 쪽에 많은 고객이 몰려있습니다. 이 자료를 기초로 평균을 구하거나, 회귀분석을 하게되면, 고객수는 적지만 많은 매출액을 가진 고객(즉, VIP)에 대한 분석에 왜곡이 생길 수 있습니다. 이런 오류를 보완하기 위해 EDA를 진행하면서 LOG변환 등을 통해 정규분포에 가깝게 변환할 것인지를 판단합니다.

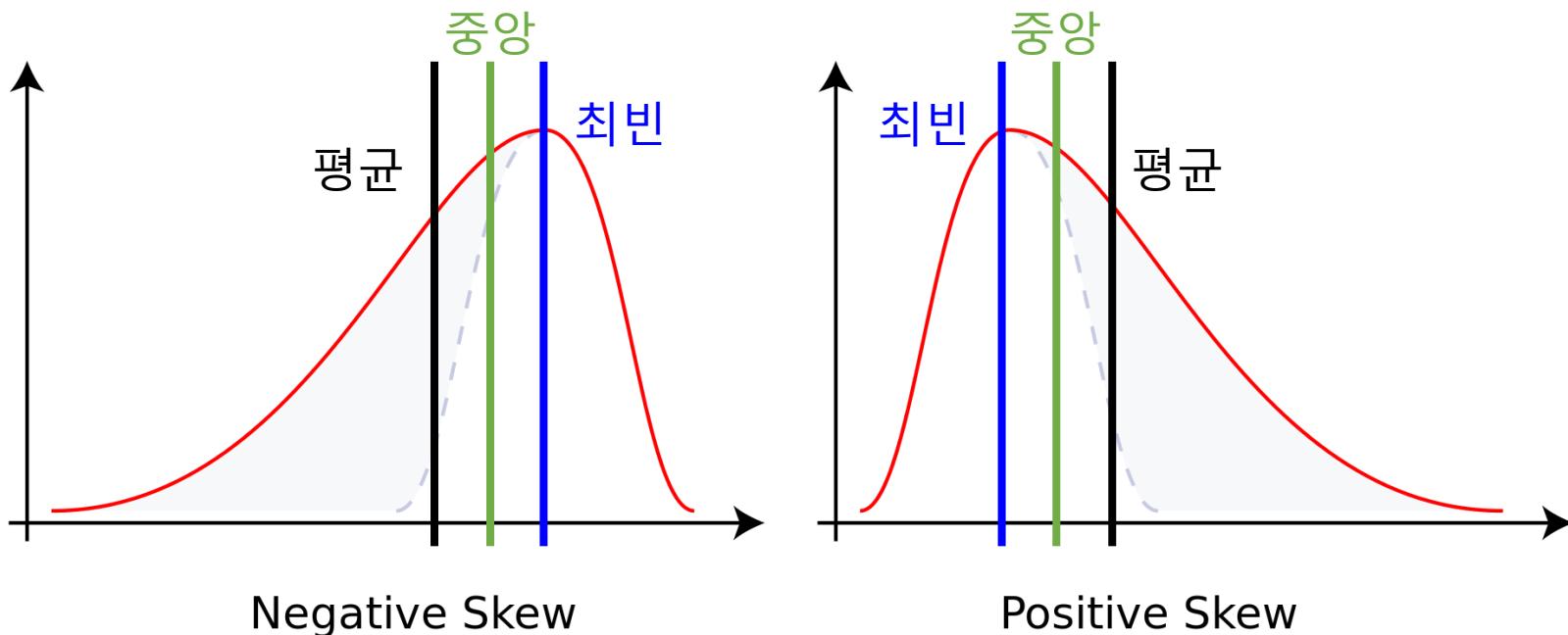
이렇게 왼쪽에 데이터가 많이 몰려서 오른쪽으로 꼬리처럼 늘어지는 형태를 skewed to the right(또는 positive skew)라고 합니다. 이 때 왜도는 양수(+)입니다. 반대로 오른쪽에 데이터가 많이 몰려서 왼쪽으로 꼬리처럼 늘어지는 형태를 skewed to the left(또는 negative skew)라고 합니다. 이 때 왜도는 음수(-)입니다.

엑셀은 조정 된 Fisher-Pearson 표준화 된 모멘트 계수(adjusted Fisher-Pearson standardized moment coefficient)를 사용합니다.

1. EDA : 왜도

27

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$



출처 : <https://ko.wikipedia.org/wiki/비대칭도>
<https://www.statisticshowto.com/skewness/>

1. EDA : 첨도

왜도가 분포의 좌우 쏠림을 표현한다면, 첨도는 분포의 꼬리가 얼마나 많이 늘어지는가를 나타내는 척도입니다.

첨도 역시 피어슨이라는 통계학자가 정리한 공식을 주로 사용하고, 이 값에서 3을 빼서 정규분포 형태를 0으로 만들어 사용하는 것이 일반적입니다.

엑셀에서는 이와는 조금 다른 공식으로 첨도를 구하지만, 기본 개념은 유사합니다.

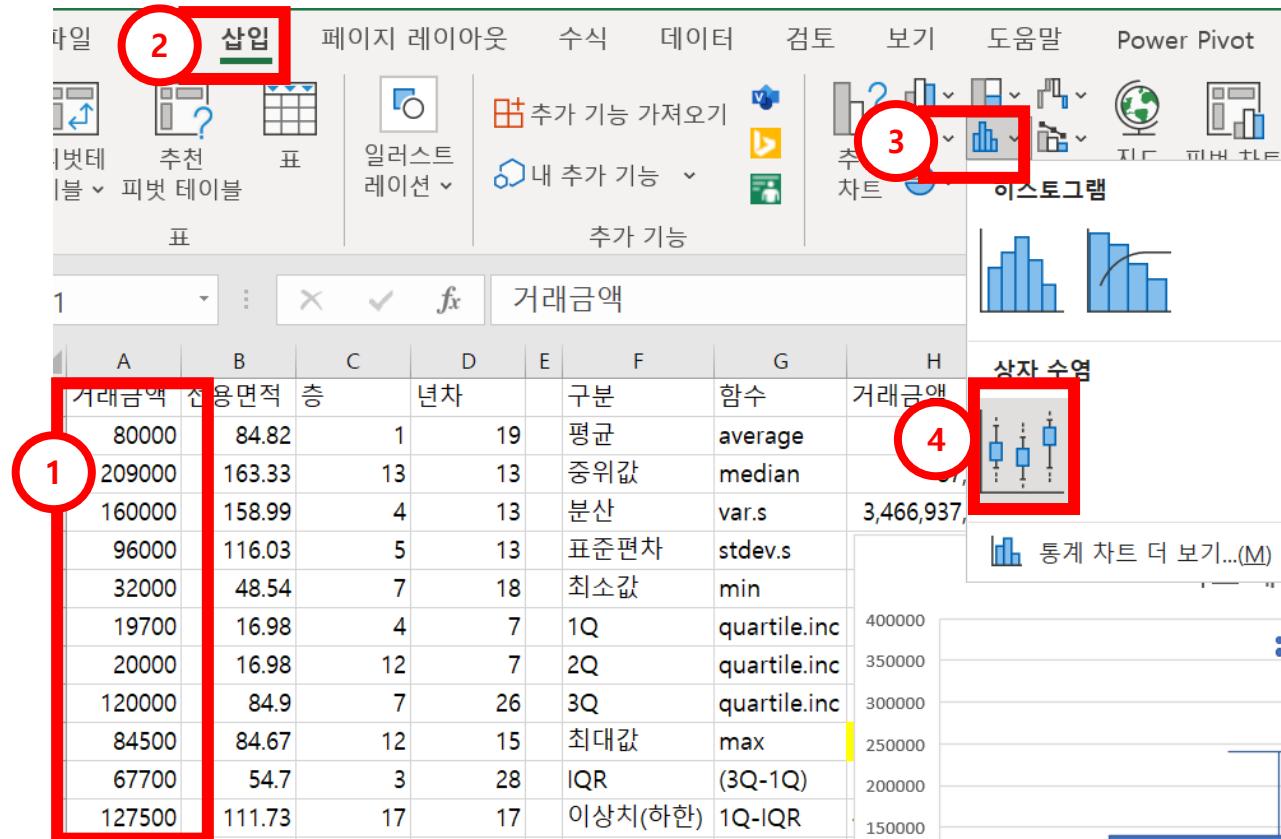
$$\left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_j - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

출처 : <https://support.microsoft.com/ko-kr/office/kurt-함수-bc3a265c-5da4-4dcb-b7fd-c237789095ab>

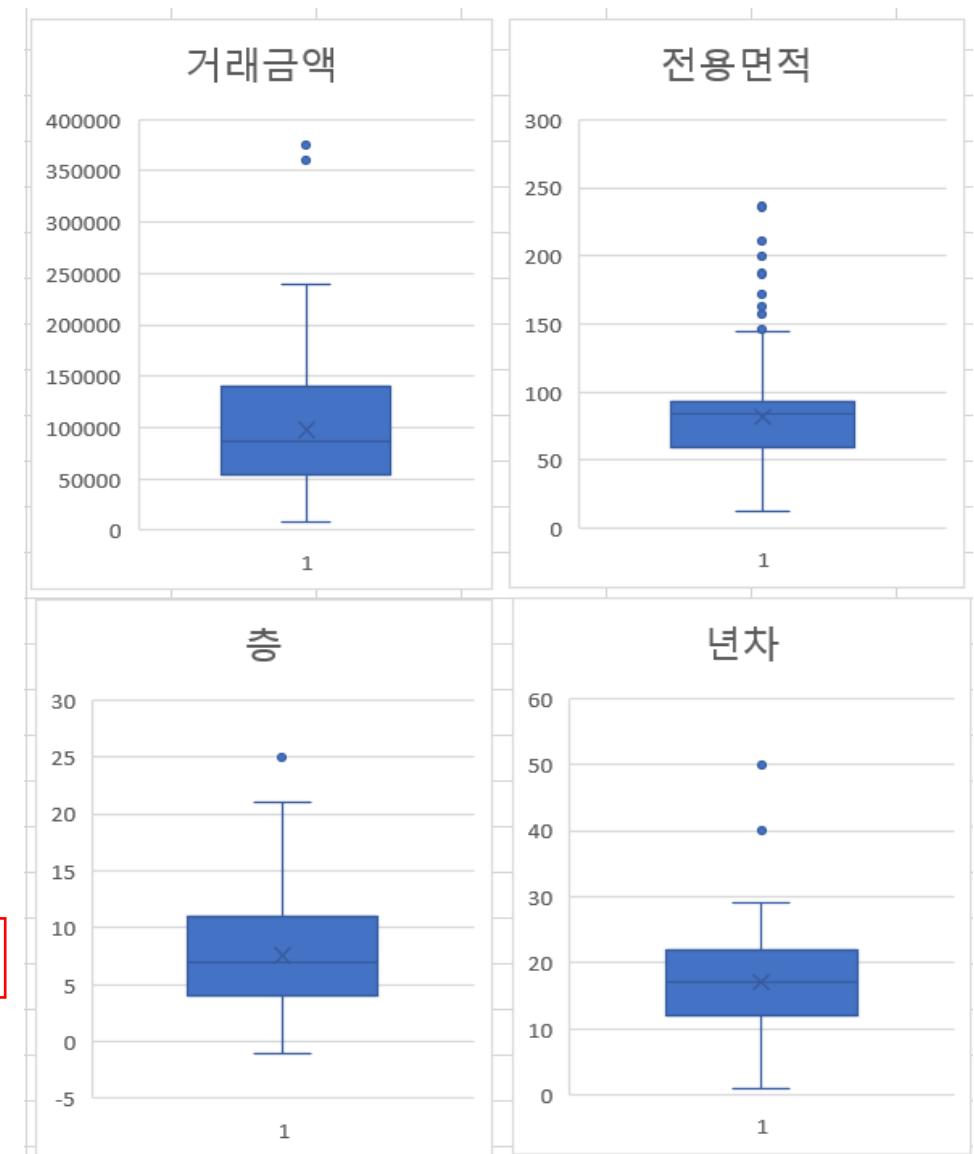
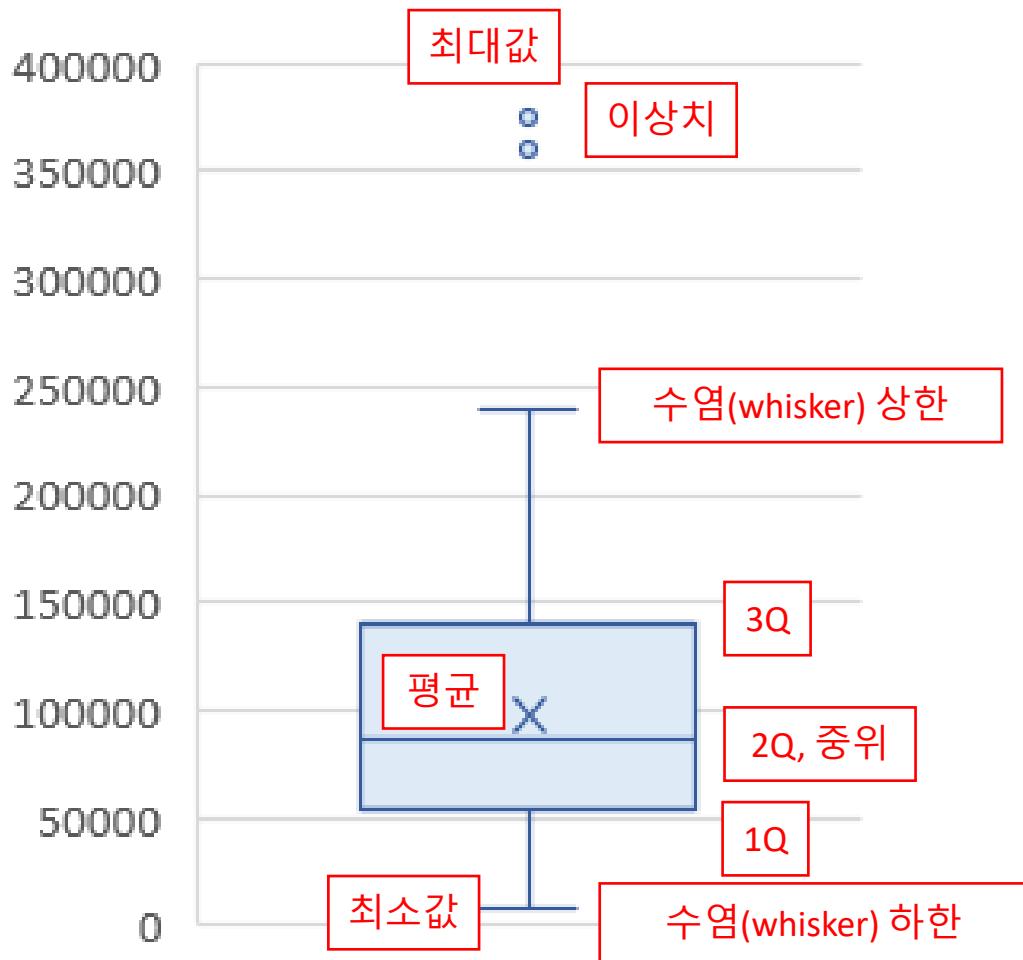
1. EDA : 박스플롯과 이상값

29

- 히스토그램을 만들 때와 같은 요령으로 이번에는 박스플롯을 그려보겠습니다.
- 그릴 열데이터를 선택한 후 삽입 > 차트 > 상자수염
- 같은 방법으로 나머지 열인 전용면적, 층, 년차도 그려봅니다.



1. EDA : 박스플롯과 이상값



1. EDA : 산점도

31

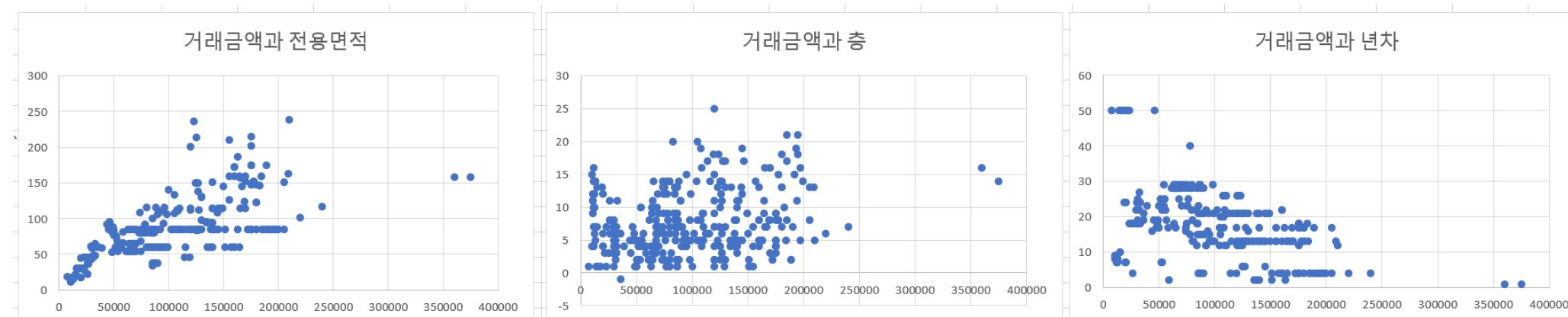
- 거래금액과 전용면적의 산점도를 그려보겠습니다.
- 두 열을 선택한 후 삽입 > 차트 > 분산형 선택

The screenshot shows the Microsoft Excel ribbon with the 'Home' tab selected (circled in red with number 2). The 'Insert' tab is also circled in red with number 2. The 'Scatter' icon in the 'Charts' section of the ribbon is circled in red with number 3. A specific scatter plot icon, which looks like a cloud of points with connecting lines, is circled in red with number 4. In the foreground, there is a data table with columns labeled '거래금액' and '전용면적'. The first row contains column headers, and the second row contains numerical values: 80000 and 84.82. The entire data range from row 1 to row 8 is highlighted with a red box. The formula bar shows '거래금액'.

	A	B	C	D	E	F	G	H
1	거래금액	전용면적	증	년차	구분	함수	거래금액	
2	80000	84.82		1	19	평균	average	97,
3	209000	163.33		13	13	중위값	median	87,
4	160000	158.99		4	13	분산	var.s	3,466,937,
5	96000	116.03		5	13	표준편차	stdev.s	58,
6	32000	48.54		7	18	최소값	min	7,
7	19700	16.98		4	7	1Q	quartile.inc	54,
8	20000	16.98		12	7	2Q	quartile.inc	87,

1. EDA : 산점도

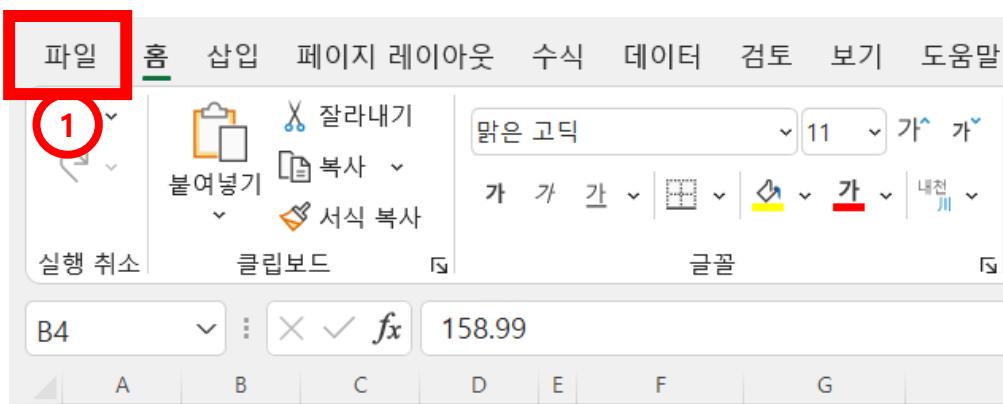
- 같은 방법으로 거래금액과 층, 거래금액과 년차의 산점도도 그려봅니다.
- 거래금액과 전용면적은 정비례 관계인 것으로 보이고, 거래금액과 층은 특별한 관계를 찾기 어렵고, 거래금액과 년차는 반비례 관계인 것으로 보입니다.
- 즉, 아파트의 면적이 클수록 거래금액이 높고, 년차가 오래되면 거래금액이 줄어든다는 것이 산점도에도 나타나고 있습니다.



1. EDA : 엑셀 추가 기능 설치

33

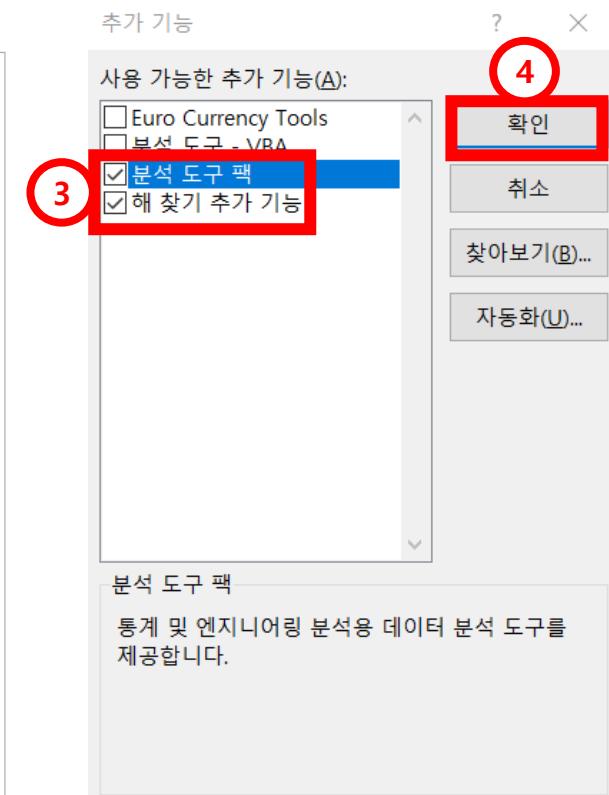
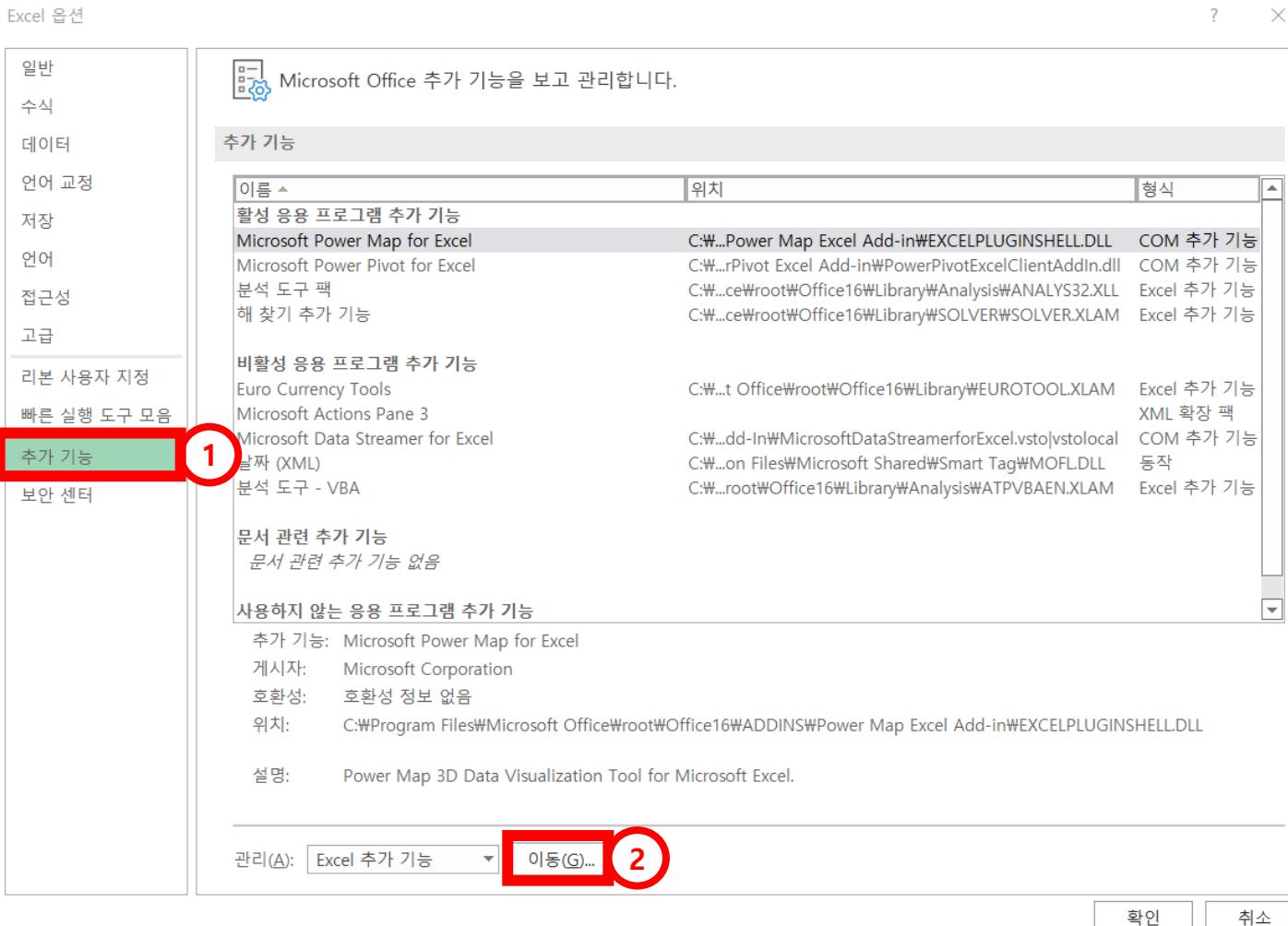
- 상단 메뉴 > 파일 > 옵션을 클릭합니다.



1. EDA : 엑셀 추가 기능 설치

34

- 'Excel 옵션' 창에서 아래쪽에 있는 '추가 기능'을 클릭한 후 '이동'버튼을 클릭하여 '추가 기능'창을 엽니다.
- '추가 기능'창이 열리면, '분석 도구 팩'과 '해 찾기 추가 기능'을 선택하고 '확인'버튼을 클릭합니다.



1. EDA : 상관분석

- 데이터 > 데이터분석 > 상관분석 > 확인

The screenshot shows a Microsoft Excel interface with the following highlights:

- 1**: The "데이터" (Data) tab is selected in the ribbon.
- 2**: The "데이터 분석" (Data Analysis) button is highlighted with a red circle.
- 3**: The "상관 분석" (Correlation) button in the "분석 도구(A)" (Analysis Tools) list is highlighted with a red circle.
- 4**: The "확인" (OK) button in the "통계 데이터 분석" (Statistical Analysis) dialog box is highlighted with a red circle.

The main worksheet displays a correlation matrix for 13 data points across various variables. A scatter plot is visible in the bottom right corner.

	A	B	C	D	E	F	G	H	I	J
1	거래금액	전용면적	층	년차	구분	함수	거래금액	전용면적	층	
2	80000	84.82		1	19	평균	average	97,886	81	8
3	209000	163.33		13	13	중위값	median	87,250	84	7
4	160000	158.99		4	13	분산	var.s	3,466,937,518	1,749	23
5	96000	116.03		5	13	표준편차	stdev.s	58,881	42	5
6	32000	48.54		7	18	최소값	min	7,500	12	1
7	19700	16.98		4	7	1Q	quartile.inc	54,000	59	4
8	20000	16.98		12	7	2Q	quartile.inc	87,250	84	7
9	120000	84.9		7	26	3Q	quartile.inc	139,800	93	11
10	84500	84.67		12	15	최대값	max	375,000	239	50
11	67700	54.7		3	28	IQR	(3Q-1Q)	128,700	50	11
12	127500	111.73		17	17	이상치(하한)	1Q-IQR	-	74,700	9
13	11600	15		7	9	이상치(상한)	3Q+IQR	268,500	143	22

1. EDA : 상관분석

36

- 입력범위를 거래금액, 전용면적, 층, 년차 열 > 첫째행 이름표 사용체크 > 새로운워크시트 > 확인
- 거래금액과 가장 높은 상관관계를 보이는 변수는 전용면적임. 0.71. 년차와는 음의상관관계 -0.41

상관 분석

입력
입력 범위(I): \$A\$1:\$D\$319

데이터 방향:
열(C) 행(R)

첫째 행 이름표 사용(1)

새로운 워크시트(P):

확인

	A	B	C	D
1	거래금액	전용면적	층	년차
2	80000	84.82	1	19
3	209000	163.33	13	13
4	160000	158.99	4	13
5	96000	116.03	5	13
6	32000	48.5	7	18
7	19700	16.93	4	7
8	20000	16.98	12	7
9	120000	84.9	7	26
10	84500	84.67	12	15
11	67700	54.7	3	28
12	127500	111.73	17	17
13	11600	15	7	9

	거래금액	전용면적	층	년차
거래금액	1.00			
전용면적	0.71	1.00		
층	0.22	-0.02	1.00	
년차	-0.41	-0.09	-0.21	1.00

1. EDA : 상관분석

37

- 입력범위를 거래금액, 전용면적, 층, 년차 열 > 첫째행 이름표 사용체크 > 새로운워크시트 > 확인
- 거래금액과 가장 높은 상관관계를 보이는 변수는 전용면적임. 0.71. 년차와는 음의상관관계 -0.41

The screenshot shows a Microsoft Excel spreadsheet with a correlation matrix. The columns and rows are labeled: 거래금액, 전용면적, 층, 년차. The values in the matrix are:

	거래금액	전용면적	층	년차
거래금액	1.00			
전용면적	0.71	1.00		
층	0.22	-0.02	1.00	
년차	-0.41	-0.09	-0.21	1.00

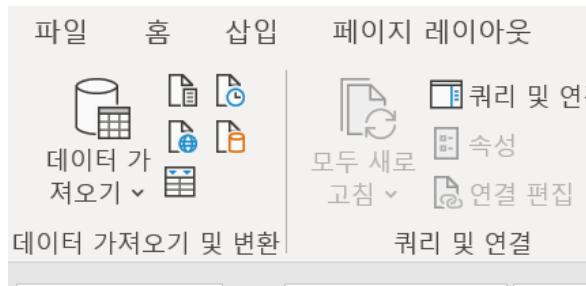
The 'Home' tab is selected in the ribbon. A context menu is open over the range B2:E5, with several items highlighted with red circles and boxes:

- 1: The first item in the context menu, '조건부 서식' (Conditional Formatting).
- 2: The 'Home' tab icon in the ribbon.
- 3: The 'Conditional Formatting' icon in the ribbon's 'Cells' section.
- 4: The '색조(S)' (Color) option under the 'Conditional Formatting' submenu.
- 5: A color swatch in the 'Color' palette.
- 6: The second item in the context menu, '셀 강조 규칙(H)' (Cell Style Rule).

1. EDA : 상관계수 계산

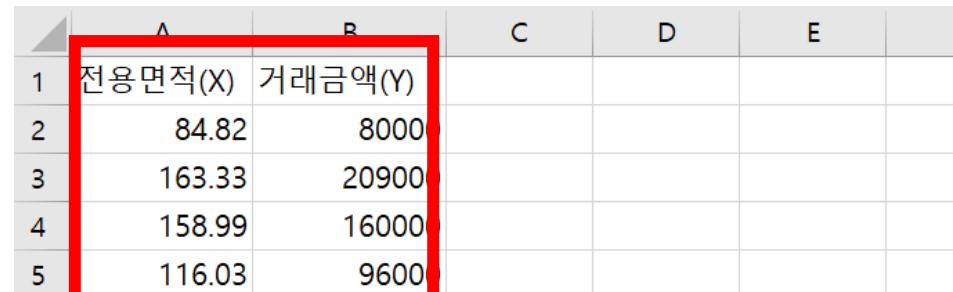
38

- 상관계수 계산을 연습하기 위해 '복사본' sheet의 데이터를 복사한 후
- 새 sheet를 만들고 '상관계수'라고 이름을 수정한 후, 상관계수 sheet에 데이터를 붙여넣습니다.
- 상관계수 계산에서 사용할 전용면적(X)과 거래금액(Y)만 남기고 모두 제거합니다.



The screenshot shows the Excel ribbon with the 'Data' tab selected. Below the ribbon, the formula bar shows 'A1' and the text '거래'. The main area displays a table with four columns: 거래금액, 전용면적, 총, and 년차. The first row contains column headers. Rows 2 through 13 show data points. A red box highlights the first seven rows (rows 2-8). A red circle with the number '1' highlights the value '16.98' in the third column of the eighth row. The entire table is enclosed in a red border.

	A	B	C	D	E
1	거래금액	전용면적	총	년차	
2	80000	84.82	1	19	
3	209000	163.33	13	13	
4	160000	158.99	4	13	
5	96000	116.03	5	13	
6	32000	48.54	7	18	
7	19700	16.98	4	7	
8	20000	16.98	12	7	
9	120000	84.9	7	26	
10	84500	84.67	12	15	
11	67700	54.7	3	28	
12	127500	111.73	17	17	
13	11600	15	7	9	



The screenshot shows the Excel ribbon with the 'Data' tab selected. The main area displays a table with two columns: '전용면적(X)' and '거래금액(Y)'. The first row contains column headers. Rows 2 through 17 show data points. A red box highlights the first seven rows (rows 2-8). A red circle with the number '3' highlights the value '84.9' in the second column of the eighth row. A red circle with the number '2' highlights the value '19700' in the second column of the eighteenth row. The entire table is enclosed in a red border. At the bottom right, the tab '상관계수' is highlighted with a red border.

	A	B
1	전용면적(X)	거래금액(Y)
2	84.82	8000
3	163.33	20900
4	158.99	16000
5	116.03	9600
6	48.54	3200
7	16.98	1970
8	16.98	2000
9	84.9	12000
10	84.67	8450
11	54.7	6770
12	111.73	12750
13		1160
14	17.811	1300
15	17.811	1250
16	17.811	1265
17	84.614	19700

1. EDA : 상관계수 계산

39

- 상관계수의 종류 : 피어슨(Pearson), 스피어맨(Spearman), 켄달(Kendall)
- 엑셀의 상관분석은 피어슨 상관계수를 사용

$$\text{피어슨상관계수} = \frac{\text{공분산}}{\text{표준편차} \cdot \text{표준편차}}$$
$$r_{XY} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{\sum_i^n (X_i - \bar{X})^2}{n-1}} \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1}}}$$

따라서

$$r_{XY} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2} \sqrt{\sum_i^n (Y_i - \bar{Y})^2}}$$

어떤 기준점에서 X, Y
만큼 움직일 때
X와 Y가 얼마나 같이
움직이는지를 구함

* 출처 : https://ko.wikipedia.org/wiki/피어슨_상관_계수

1. EDA : 상관계수 계산

- 1번 영역에서는 분자를 구하고, 2번 영역에서는 분모를 구하여 상관계수를 계산합니다.

	A	B	C	D	E	F	G	I	J
1	전용면적(X)	거래금액(Y)	X-X평균	Y-Y평균	C*D	C^2	D^2	N	318
2	84.82	80,000	3	- 17,886	- 59,521	11	319,926,095	전용면적(X) 평균	81.5
3	163.33	209,000	82	111,114	9,093,276	6,697	12,346,214,774	거래금액(Y) 평균	97,886.5
4	158.99	160,000	77	62,114	4,813,656	6,006	3,858,089,617		
5	116.03	96,000	35	- 1,886	- 65,155	1,193	3,558,799	부자	552,012,548
6	48.54	32,000	- 33	- 65,926	- 2,171,110	1,096	4,241,027,982		
7	16.98	19,700	- 65	-	- 0,043,989	4,111	5,340	SUM(F)	554,388
8	16.98	20,000	- 65	-	- 0,024,635	4,111	3,453	SUM(G)	1,099,019,193,055
9	84.9	120,000	3	-	75,357	7,856	7,856	SQRT(SUM(F))	745
10	84.67	84,500	3	- 13,386	- 42,538	10	179,197,793	SQRT(SUM(G))	1,048,341
11	54.7	67,700	- 27	- 30,186	808,765	718	911,223,453	분모	780,566,137
12	111.73	127,500	30	29,614	895,445	914	876,960,686		
13	15	11,600	- 66	- 86,286	5,737,385	4,421	7,445,356,283	상관계수	0.71
14	17.811	13,000	- 64	- 84,886	5,405,680	4,055	7,205,714,145		
15	17.811	12,500	- 64	- 85,386	5,437,521	4,055	7,290,850,623		
16	17.811	12,650	- 64	- 85,336	5,437,020	4,055	7,295,357,100		

1. EDA : 상관계수 계산

- 공분산을 구하는 식에서 $n-1$ 을 약분한 값인 분자 부분을 먼저 구하겠습니다.
- (1)전체 데이터의 갯수를 구하는 N값을 만듭니다. J2에 `=COUNT(B2:B319)`을 입력합니다.
- J3과 J4에 각각 X와 Y의 평균을 구하는 식을 입력합니다. J3에는 `=AVERAGE(A2:A319)`를 J4에는 `=AVERAGE(B2:B319)`를 입력합니다.
- (2)C열에 $X-\bar{X}$ 평균값을 구하기 위해 C2에 `=A2-\$J\$2`를 입력한 후 아래행에도 수식을 복사해서 값을 채워줍니다.
- (3)D열에 $Y-\bar{Y}$ 평균값을 구하기 위해 D2에 `=B2-\$J\$3`를 입력한 후 아래행에도 수식을 복사해서 값을 채워줍니다.
- (4) $(X-\bar{X})^*(Y-\bar{Y})$ 을 구하기 위해 E2에 `=C2*D2`를 입력한 후 아래행에도 수식을 복사해서 값을 채워줍니다.
- (5)4번에서 구한 값을 전부 합해줍니다. J5에 `=SUM(E2:E319)`를 입력하여 분자값을 완성합니다.

1. EDA : 상관계수 계산

$$r_{XY} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2} \sqrt{\sum_i^n (Y_i - \bar{Y})^2}}$$

	A	B	C	D	E	F	G
1	전용면적(X)	거래금액(Y)	X-X평균	Y-Y평균	C*D	C^2	D^2
2	84.82	80,000	3	17,886	-59,521	11	319,926,095
3	163.33	209,000	82	111,114	9,093,276	6,697	12,346,214,774
4	158.99	160,000	77	62,114	4,813,656	6,006	3,858,089,617
5	116.03	96,000	35	1,886	-65,155	1,193	3,558,799
6	48.54	32,000	-33	65,886	2,171,110	1,086	4,341,027,982
7	16.98	19,700	-65	186	5,00989	4,162	6,113,125,340
8	16.98	20,000	-2	36	5,0035	4,162	6,066,303,453
9	84.9	120,000	3	22,114	75,357	12	489,007,856
10	84.67	84,500	3	13,386	-42,538	10	179,197,793
11	54.7	67,700	-27	30,186	808,765	718	911,223,453
12	111.73	127,500	30	29,614	895,445	914	876,960,686
13	15	11,600	-66	86,286	5,737,385	4,421	7,445,356,283
14	17.811	13,000	-64	84,886	5,405,680	4,055	7,205,714,145
15	17.811	12,500	-64	85,386	5,437,521	4,055	7,290,850,623
16	17.811	12,650	-64	85,236	5,427,960	4,055	7,265,257,180

I	1	J
N	318	
전용면적(X) 평균	81.5	
거래금액(Y) 평균	97,886.5	
분자	5	552,012,548
SUM(F)		554,388
SUM(G)		1,099,019,193,055
SQRT(SUM(F))		745
SQRT(SUM(G))		1,048,341
분모		780,566,137
상관계수		0.71

1. EDA : 상관계수 계산

- 공분산을 구하는 식에서 $n-1$ 을 약분한 값인 분자 부분을 먼저 구하겠습니다.

$$r_{XY} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2} \sqrt{\sum_i^n (Y_i - \bar{Y})^2}}$$

The diagram illustrates the formula for calculating the correlation coefficient r_{XY} . The formula is shown as:

$$r_{XY} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2} \sqrt{\sum_i^n (Y_i - \bar{Y})^2}}$$

The components are highlighted with colored boxes and numbered circles:

- Step 1:** The term $(X_i - \bar{X})(Y_i - \bar{Y})$ is highlighted in red.
- Step 2:** The term $\sqrt{\sum_i^n (X_i - \bar{X})^2}$ is highlighted in green.
- Step 3:** The term $\sqrt{\sum_i^n (Y_i - \bar{Y})^2}$ is highlighted in blue.
- Step 4:** The entire denominator $\sqrt{\sum_i^n (X_i - \bar{X})^2} \sqrt{\sum_i^n (Y_i - \bar{Y})^2}$ is highlighted in black.
- Step 5:** The overall fraction is highlighted in purple.
- Step 6:** The label r_{XY} is highlighted in orange.

	A	B	C	D	E	F	G	
1	전용면적(X)	거래금액(Y)	X-X평균	Y-Y평균	C*D	C^2	D^2	
2	84.82	80,000	3	- 17,886	- 59,521	11	319,926,095	
3	163.33	209,000	82	111,114	9,093,276	6,697	12,346,214,774	
4	158.99	160,000	77	62,114	4,813,656	6,006	3,858,089,617	
5	116.03	96,000	35	- 1,886	- 65,155	1,193	3,558,799	
6	48.54	32,000	-	33	65,886	2,171,110	1,086	4,341,027,982
7	16.98	19,700	-	65	- 78,186	5,043,989	162	6,112,25,340
8	16.98	20,000	-	65	- 77,886	5,024,635	162	6,062,0,453
9	84.9	120,000	3	22,114	75,357	12	489,007,856	
10	84.67	84,500	3	- 13,386	- 42,538	10	179,197,793	
11	54.7	67,700	-	27	- 30,186	808,765	718	911,223,453
12	111.73	127,500	30	29,614	895,445	914	876,960,686	
13	15	11,600	-	66	- 86,286	5,737,385	4,421	7,445,356,283
14	17.811	13,000	-	64	- 84,886	5,405,680	4,055	7,205,714,145
15	17.811	12,500	-	64	- 85,386	5,437,521	4,055	7,290,850,623
16	17.811	12,650	-	64	- 85,226	5,427,969	4,055	7,265,257,180

I	J
N	318
전용면적(X) 평균	81.5
거래금액(Y) 평균	97,886.5
분자	552,012,548
SUM(F)	554,388
SUM(G)	1,099,019,193,055
SQRT(SUM(F))	745
SQRT(SUM(G))	1,048,341
분모	780,566,137
상관계수	0.71

1. EDA : 상관계수 계산

	A	B	C	D	E	F	G
1	전용면적(X)	거래금액(Y)	X-X평균	Y-Y평균	C*D	C^2	D^2
2	84.82	80000	=A2-\$J\$2	=B2-\$J\$3	=C2*D2	=C2^2	=D2^2
3	163.33	209000	=A3-\$J\$2	=B3-\$J\$3	=C3*D3	=C3^2	=D3^2
4	158.99	160000	=A4-\$J\$2	=B4-\$J\$3	=C4*D4	=C4^2	=D4^2
5	116.03	96000	=A5-\$J\$2	=B5-\$J\$3	=C5*D5	=C5^2	=D5^2
6	48.54	32000	=A6-\$J\$2	=B6-\$J\$3	=C6*D6	=C6^2	=D6^2

	H	I	J
1		N	=COUNT(B2:B319)
2		전용면적(X) 평균	=AVERAGE(A2:A319)
3		거래금액(Y) 평균	=AVERAGE(B2:B319)
4			
5		분자	=SUM(E2:E319)
6			
7		SUM(F)	=SUM(F2:F319)
8		SUM(G)	=SUM(G2:G319)
9		SQRT(SUM(F))	=SQRT(J7)
10		SQRT(SUM(G))	=SQRT(J8)
11		분모	=J9*J10
12			
13		상관계수	=J5/J11
14			
15			

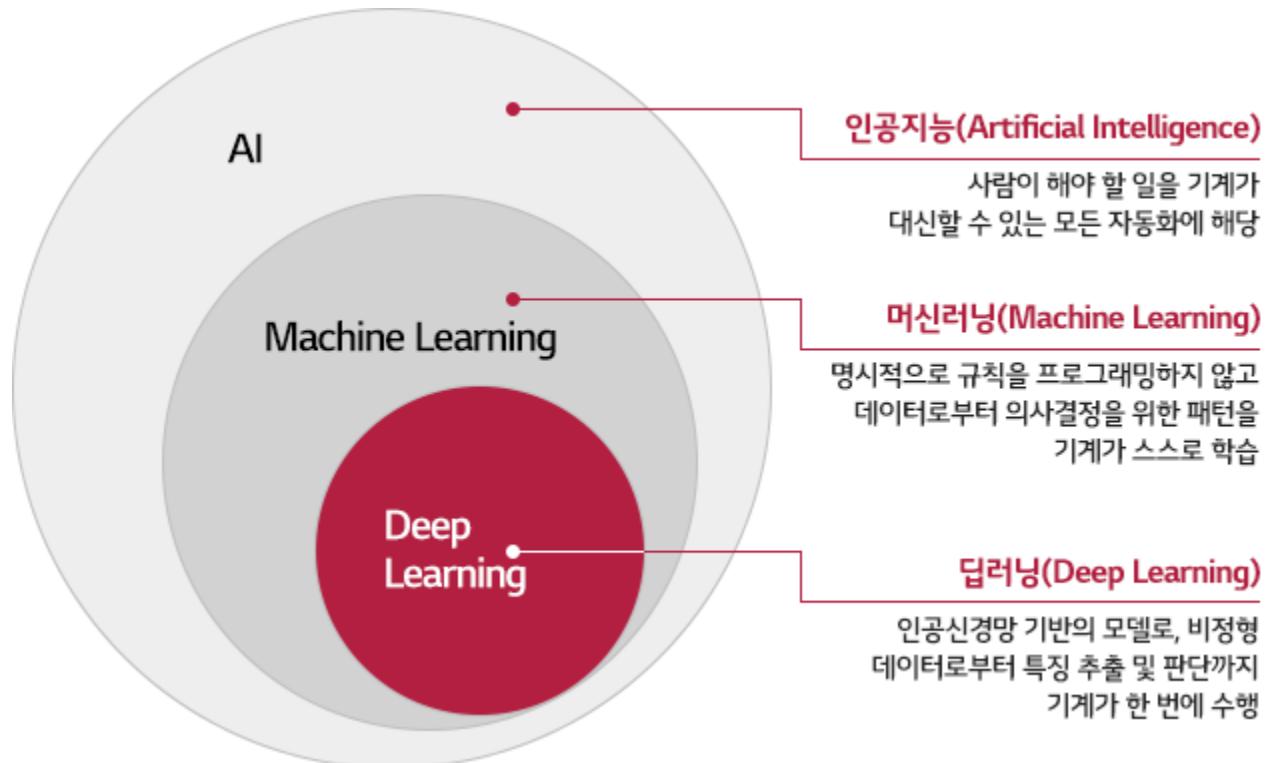
1. EDA : 상관계수 계산

- 엑셀에서는 함수로 상관계수를 구할 수 있습니다.
- (1)CORREL함수를 이용하여 바로 구할 수 있습니다. `=CORREL(A2:A319,B2:B319)`를 입력하여 구합니다.
- (2) 공분산 / (X의 표준편차 * Y의 표준편차)를 계산하여 구할 수 있습니다.
- * 공분산 : COVARIANCE.S함수 사용(`=COVARIANCE.S(A2:A319,B2:B319)`)
- * X의 표준편차 : VAR.S함수 사용(`=VAR.S(A2:A319)`)
- * Y의 표준편차 : VAR.S함수 사용(`=VAR.S(B2:B319)`)

CORREL	0.71	①	CORREL	=CORREL(A2:A319,B2:B319)
COVARIANCE.S	1,741,365		COVARIANCE.S	=COVARIANCE.S(A2:A319,B2:B319)
X의 VAR.S	1,749	②	X의 VAR.S	=VAR.S(A2:A319)
Y의 VAR.S	3,466,937,518		Y의 VAR.S	=VAR.S(B2:B319)
COV/SQRT(VAR_X*VAR_Y)	0.71		COV/SQRT(VAR_X*VAR_Y)	=M4/(SQRT(M6)*SQRT(M7))

2. 비지도학습

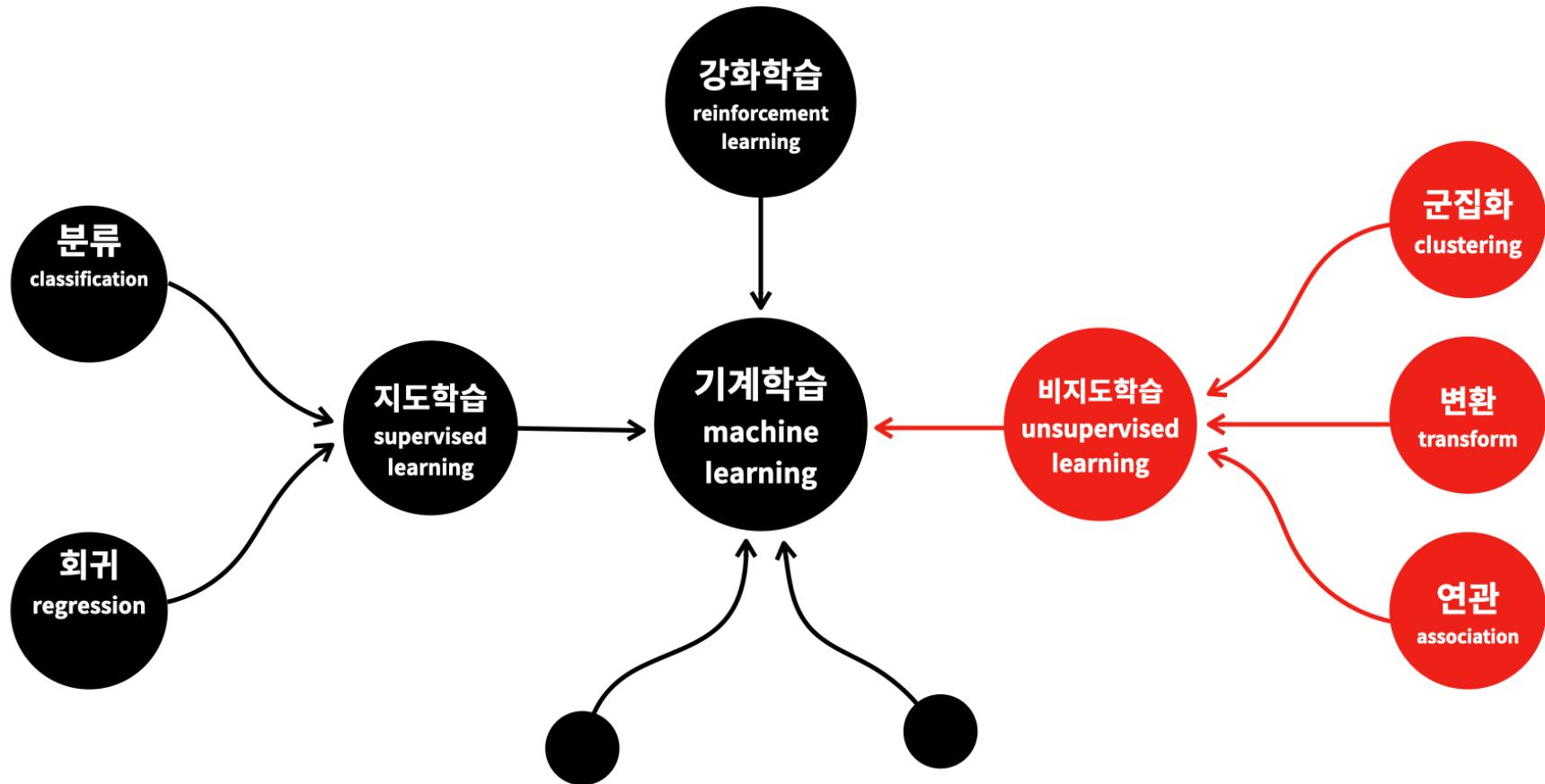
46



cf) 통계

2. 비지도학습

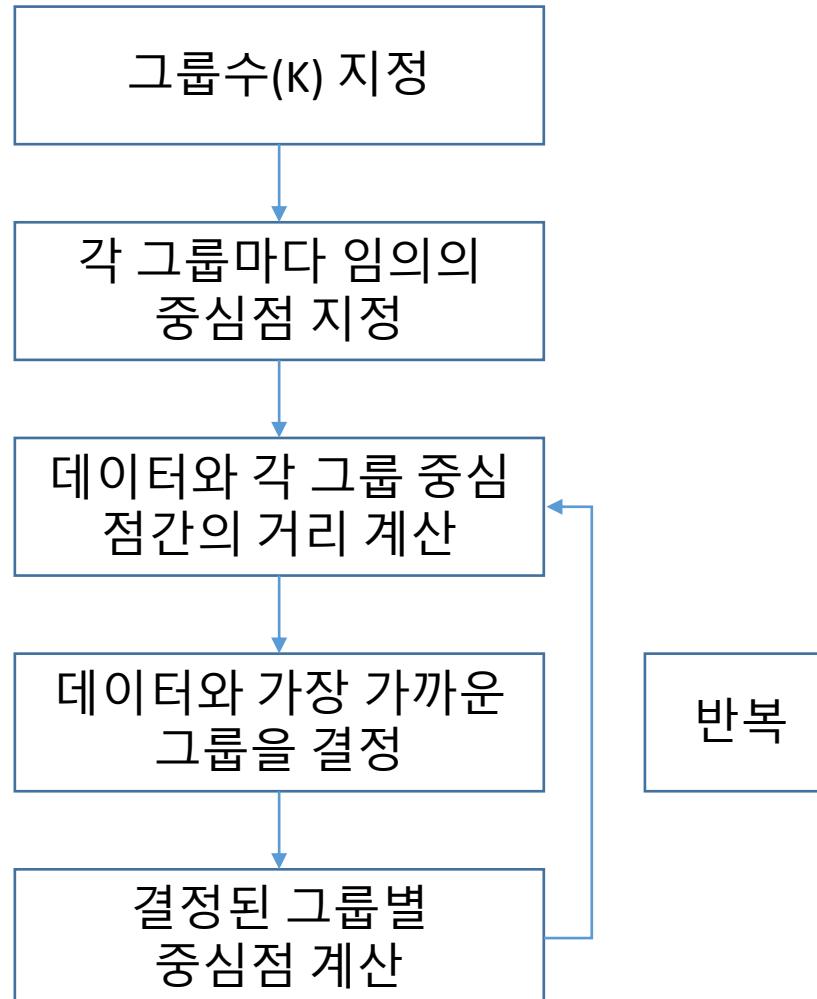
47



2. 비지도학습 : K-means

48

- K-means 알고리즘 계산 순서



2. 비지도학습 : K-means

- 데이터를 K개의 군집으로 나누어주는 대표적인 비지도학습 방법인 k-means를 계산을 통해 이해
 - 실거래 데이터 중 층과 연차 데이터를 사용하겠습니다.
 - A그룹을 층 2, 연차 20으로 가정하고 / B그룹을 층 15, 연차 10으로 가정한 상황입니다.(임의의 수치)

2. 비지도학습 : K-means

- A의 중심점인 (2,20)으로 부터 얼마나 떨어져있는지 유클리디안 거리를 계산합니다.
- 우선 (1)D3, E3셀에 2와 20을 입력합니다. 아래에 있는 (2)D4셀에 `=(A4-D\$3)^2`, (3)E4셀에 `=(B4-E\$3)^2`을 입력합니다.
- 2개 열에 대한 차이(거리)의 제곱을 구했으니, 이제 2개 값을 합한 후 제곱근을 구하겠습니다. (4)F4셀에 `=SQRT(D4+E4)`를 입력하여 D4와 E4에서 구한 값을 합하고 제곱근을 씌워줍니다.

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad \text{유클리디안 거리}$$

	A	B	C	D	E	F	G	H	I	J
1	증과 년차로 이루어진 2차원 데이터를 A그룹과 B그룹 2개(k)의 그룹으로 나눕니다.									
2	[A를 2, 20으로 가정]					[B를 15와 10으로 가정]				
3	증	년차		2	20	거리	15	10	거리	A-B
4	1	19		1	1		196	81	17	- 15
5	13	13		121	49		4	9	4	9
6	4	13		4	49		121	9	11	- 4
7	5	13		9	49		100	9	10	- 3

	A	B	C	D	E	F	G	H	I	J
1	증과 년차로 0									
2	[A를 2, 20으로 가정]					[B를 15와 10으로 가정]				
3	증	년차	1	2	20	거리	15	10	거리	A-B
4	1	19		=(\$A4-D\$3)^2	=(\$B4-E\$3)^2	=SQRT(D4+E4)	=(\$A4-G\$3)^2	=(\$B4-H\$3)^2	=SQRT(G4+H4)	=F4-I4
5	13	13		=(\$A5-D\$3)^2	=(\$B5-E\$3)^2	=SQRT(D5+E5)	=(\$A5-G\$3)^2	=(\$B5-H\$3)^2	=SQRT(G5+H5)	=F5-I5
6	4	13		=(\$A6-D\$3)^2	=(\$B6-E\$3)^2	=SQRT(D6+E6)	=(\$A6-G\$3)^2	=(\$B6-H\$3)^2	=SQRT(G6+H6)	=F6-I6

3

4

2. 비지도학습 : K-means

- 이번에는 같은 4행의 값(1, 19)가 B의 중심점인 (15, 10)으로 부터 얼마나 떨어져있는지 유clidean 거리를 계산합니다. (1)G3, H3셀에 B그룹의 기준값인 1과 19를 입력합니다. 아래에 있는 (2)G4셀에 `=(A4-G\$3)^2`, (3)H4셀에 `=(B4-H\$3)^2`을 입력합니다. 2개 열에 대한 차이(거리)의 제곱을 구했으니, 이 2개 값을 합한 후 제곱근을 구하겠습니다. (4)I4셀에 `=SQRT(G4+H4)`를 입력하여 G4와 H4에서 구한 값을 합하고 제곱근을 씌워줍니다.
- 마지막으로 (5)J4셀에 `=F4-I4`를 입력하여 A그룹에 속할 것인지 B그룹에 속할 것인지 판단할 수 있게합니다.(음수가 나오면 A그룹과의 거리가 더 가까운 것이므로 A그룹이라 지정하고, 양수가 나오면 그 반대이므로 B그룹이라 정합니다)

	A	B	C	D	E	F	G	H	I	J
1	증과 년차로 이루어진 2차원 데이터를 A그룹과 B그룹 2개(k)의 그룹으로 나눕니다.									
2						[A를 2, 20으로 가정]		[B를 15와 10으로 가정]		
3	증	년차		2	20	거리	1	15	10	거리
4	1	19		1	1	1	196	81	17	A-B
5	13	13		121	49	13	4	9	4	15
6	4	13		4	49	7	121	9	11	9
7	5	13		9	49	8	100	9	10	4

	A	B	C	D	E	F	G	H	I	J
1	증과 년차로 0									
2						[A를 2, 20으로 가정]			[B를 15와 10으로 가정]	
3	증	년차	2	20	거리	1	15	10	거리	A-B
4	1	19	=(\$A4-D\$3)^2	=(\$B4-E\$3)^2	=SQRT(D4+E4)	2	=(\$A4-G\$3)^2	=(\$B4-H\$3)^2	=SQRT(G4+H4)	=F4-I4
5	13	13	=(\$A5-D\$3)^2	=(\$B5-E\$3)^2	=SQRT(D5+E5)	=(\$A5-G\$3)^2	=(\$B5-H\$3)^2	=SQRT(G5+H5)	=F5-I5	
6	4	13	=(\$A6-D\$3)^2	=(\$B6-E\$3)^2	=SQRT(D6+E6)	=(\$A6-G\$3)^2	=(\$B6-H\$3)^2	=SQRT(G6+H6)	=F6-I6	

2. 비지도학습 : K-means

52

- 작업을 편리하게 하기 위해 윗쪽 3개행과 왼쪽 3개열을 기준으로 틀고정하겠습니다. (1)D4셀을 선택한 후 (2)보기 > 틀 고정 > 틀 고정을 선택합니다.

1 총과 년차로 이루어진 2차원 데이터를 A그룹과 B그룹 2개(k)의 그룹으로 나눕니다.

2 [A를 2, 20으로 가정] [B를 15와 10으로 가정]

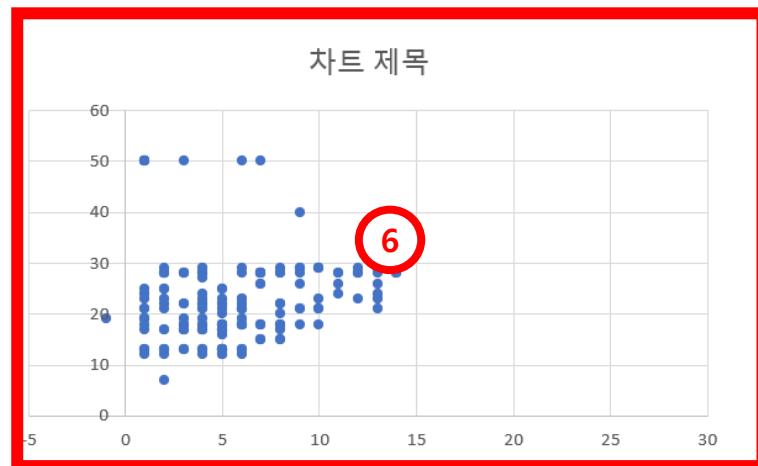
	총	년차	2	20	거리	15	10	거리	A-B
4	1	19	1	1	196	81	17	-	1
5	13	13	121	49	13	4	9	4	9
6	4	13	4	49	7	121	9	11	4
7	5	13	9	49	8	100	9	10	3

2. 비지도학습 : K-means

- 산점도를 그려보기 위해 A그룹에 속한 년차와 B그룹에 속한 년차를 별도로 계산합니다.
- 먼저 층과 A그룹을 선택한 후 산점도를 그립니다.

The screenshot shows the Microsoft Excel ribbon with the '삽입' (Insert) tab highlighted (circled in red). In the 'Insert' tab's ribbon group, the 'Scatter' icon is circled in red. Below the ribbon, a data table is displayed with columns labeled '총', '년차', 'A년차', and 'B년차'. The 'A년차' column contains values 19, 13, 13, 13, 18, #N/A, 7, 26, 15, and 28. The 'B년차' column contains values #N/A, 13, #N/A, #N/A, #N/A, 7, 7, #N/A, 15, and #N/A. The first two rows of the 'A년차' column are also circled in red.

K	L
A년차	B년차
=IF(J4<=0,\$B4,NA())	=IF(J4>0,\$B4,NA())
=IF(J5<=0,\$B5,NA())	=IF(J5>0,\$B5,NA())
=IF(J6<=0,\$B6,NA())	=IF(J6>0,\$B6,NA())
=IF(J7<=0,\$B7,NA())	=IF(J7>0,\$B7,NA())
=IF(J8<=0,\$B8,NA())	
=IF(J9<=0,\$B9,NA())	



2. 비지도학습 : K-means

54

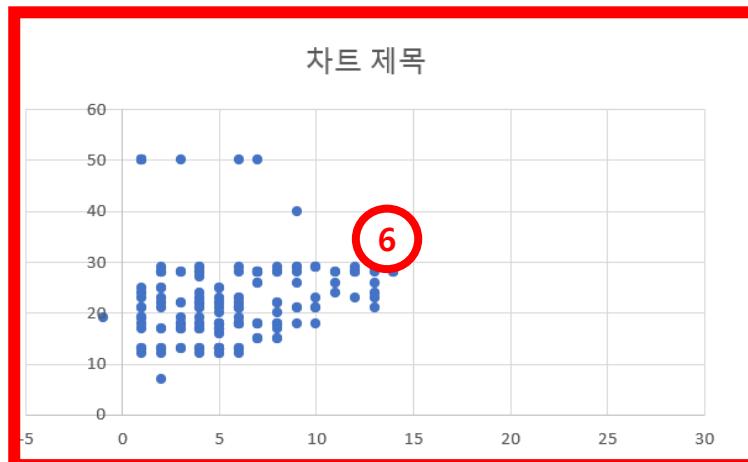
- 이번에는 층과 B그룹을 선택한 후 B그룹에 대한 산점도를 그립니다.

The screenshot shows the Microsoft Excel ribbon with the '삽입' (Insert) tab highlighted (step 3). In the chart tools ribbon, the '분산형' (Scatter) icon is selected (step 4). A scatter plot is displayed on the worksheet (step 5), showing data points for two groups. The worksheet contains the following data:

	A	B	C	K	L
1	층과 년차로 이동				
2					
3	층	년차	A년차	B년차	
4	1	19	19	#N/A	
5	2	13	#N/A	2	
6	4	13	13	#N/A	
7	5	13	13	#N/A	
8	7	18	18	#N/A	
9	4	7	#N/A	7	
10	12	7	#N/A	7	
11	7	26	26	#N/A	
12	12	15	#N/A	15	
13	3	28	28	#N/A	

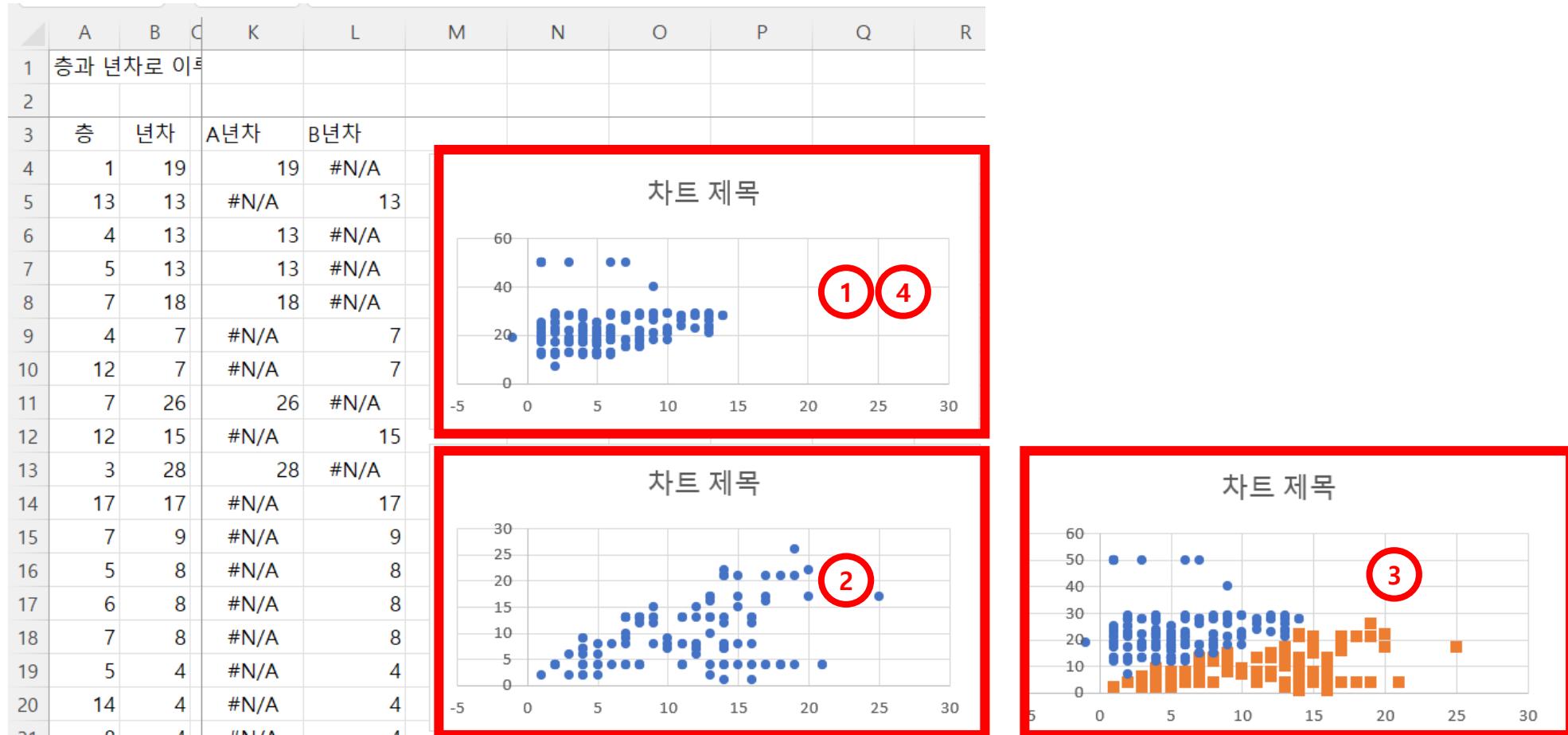
The formula bar displays the following conditional logic for the B년차 column:

K	L
A년차	B년차
=IF(J4<=0,\$B4,NA())	=IF(J4>0,\$B4,NA())
=IF(J5<=0,\$B5,NA())	=IF(J5>0,\$B5,NA())
=IF(J6<=0,\$B6,NA())	=IF(J6>0,\$B6,NA())
=IF(J7<=0,\$B7,NA())	=IF(J7>0,\$B7,NA())



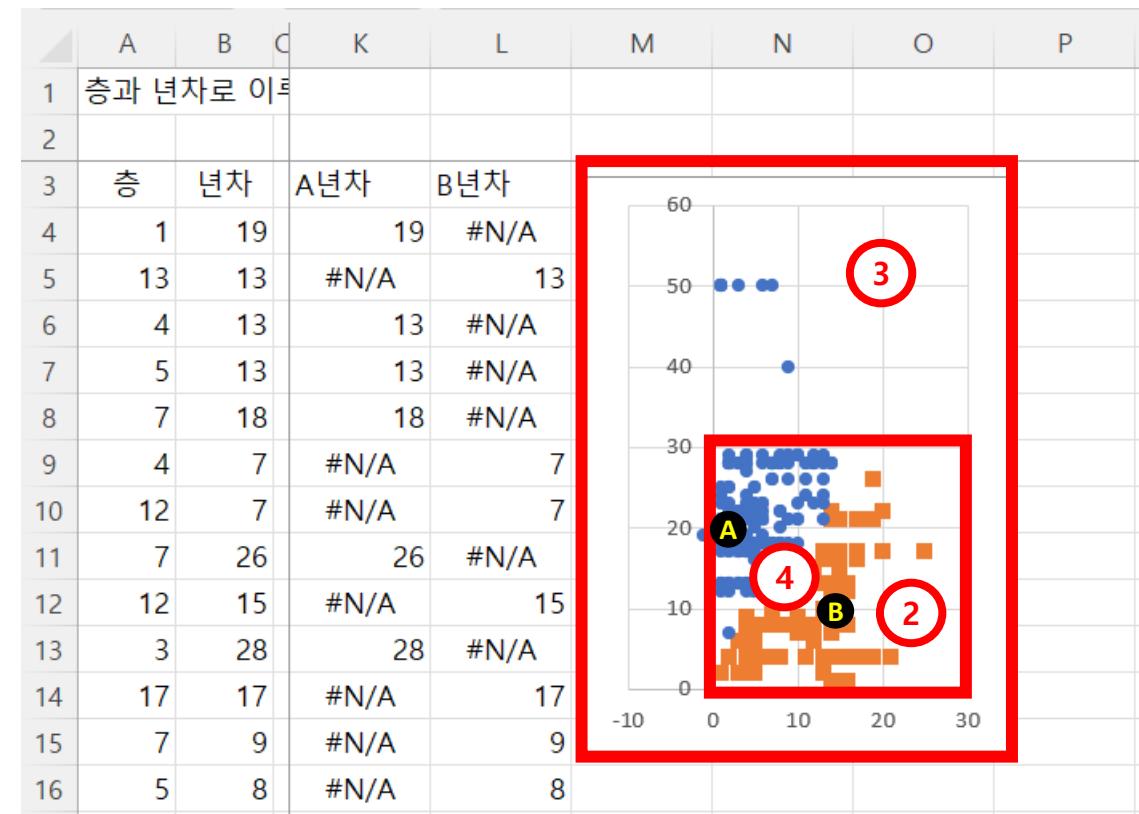
2. 비지도학습 : K-means

- 완성한 (1)A그룹의 산점도를 복사한 후 (2)B그룹의 산점도를 선택하고 붙여넣기 합니다.
- (3)통합된 차트만 남기고 A그룹의 산점도를 (4)삭제합니다.



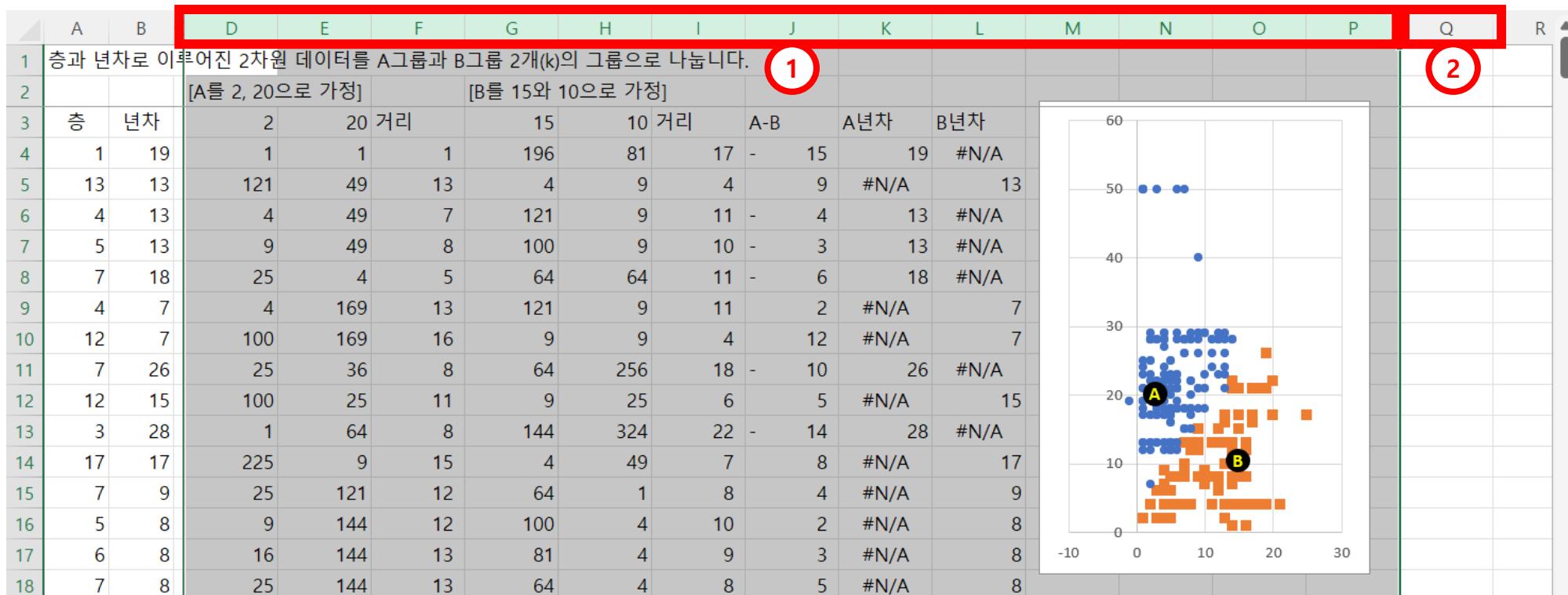
2. 비지도학습 : K-means

- (1) 차트 제목을 삭제합니다.
- (2) X축과 Y축의 스케일이 서로 다르기 때문에 같은 숫자인 30을 기준으로 볼 때 서로 다른 거리로 표현되고 있습니다.
- (3) 차트 크기를 조정하여 X축과 Y축이 동일한 거리를 나타낼 수 있도록 수정합니다.



2. 비지도학습 : K-means

- 이제 2번째 단계 계산을 위해 D열부터 P열을 선택한 후 복사합니다.
- Q열을 선택한 후 붙여넣기하면, 새로운 단계가 생성됩니다.



2. 비지도학습 : K-means

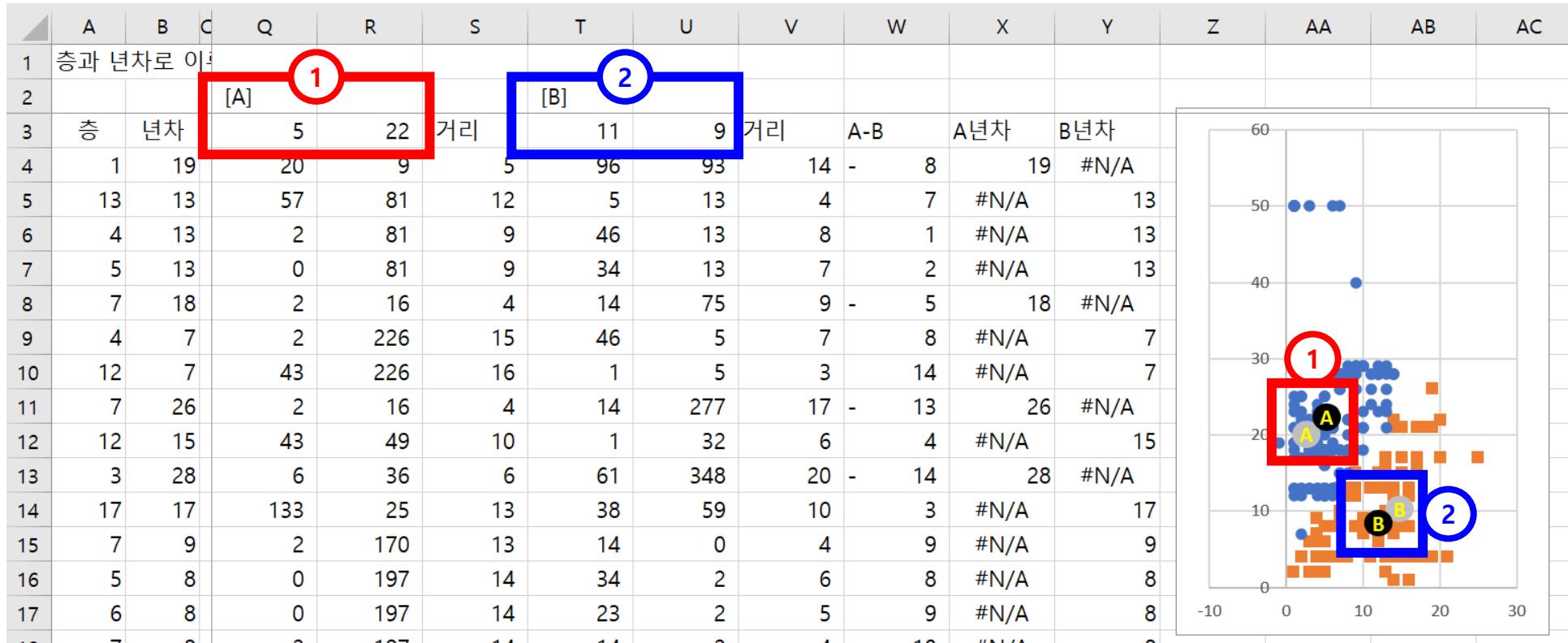
- 1단계에서는 추정치로 중심점을 잡았지만 1차 계산을 했기 때문에 이제는 실제 A그룹과 B그룹의 중심점을 구해서 2단계 계산에 적용해보겠습니다.
- Averageif 함수를 사용하여 A그룹의 값에 해당하는 층과 년차의 평균을 구하고, B그룹에 해당하는 층과 년차를 구하겠습니다.
- Averageif는 조건이 되는 범위, 조건, 평균을 구할 범위로 구성되어 있습니다.
- 여기서는 K열에 A그룹이면 년차를 표시하도록 되어 있기 때문에, A그룹에 대해서만 평균을 구하고 싶다면, 조건이 되는 범위를 K열로 하고, 조건을 ">=0" 0이상으로 지정합니다.
- 평균을 구할 범위는 각각 층과, 년차에 해당하는 A열과 B열을 선택해주시면 됩니다.

Q	R
[A] =AVERAGEIF(\$K\$4:\$K\$321, ">=0",\$A\$4:\$A\$321)	=AVERAGEIF(\$K\$4:\$K\$321, ">=0",\$B\$4:\$B\$321)

T	U
[B] =AVERAGEIF(\$L\$4:\$L\$321, ">=0",\$A\$4:\$A\$321)	=AVERAGEIF(\$L\$4:\$L\$321, ">=0",\$B\$4:\$B\$321)

2. 비지도학습 : K-means

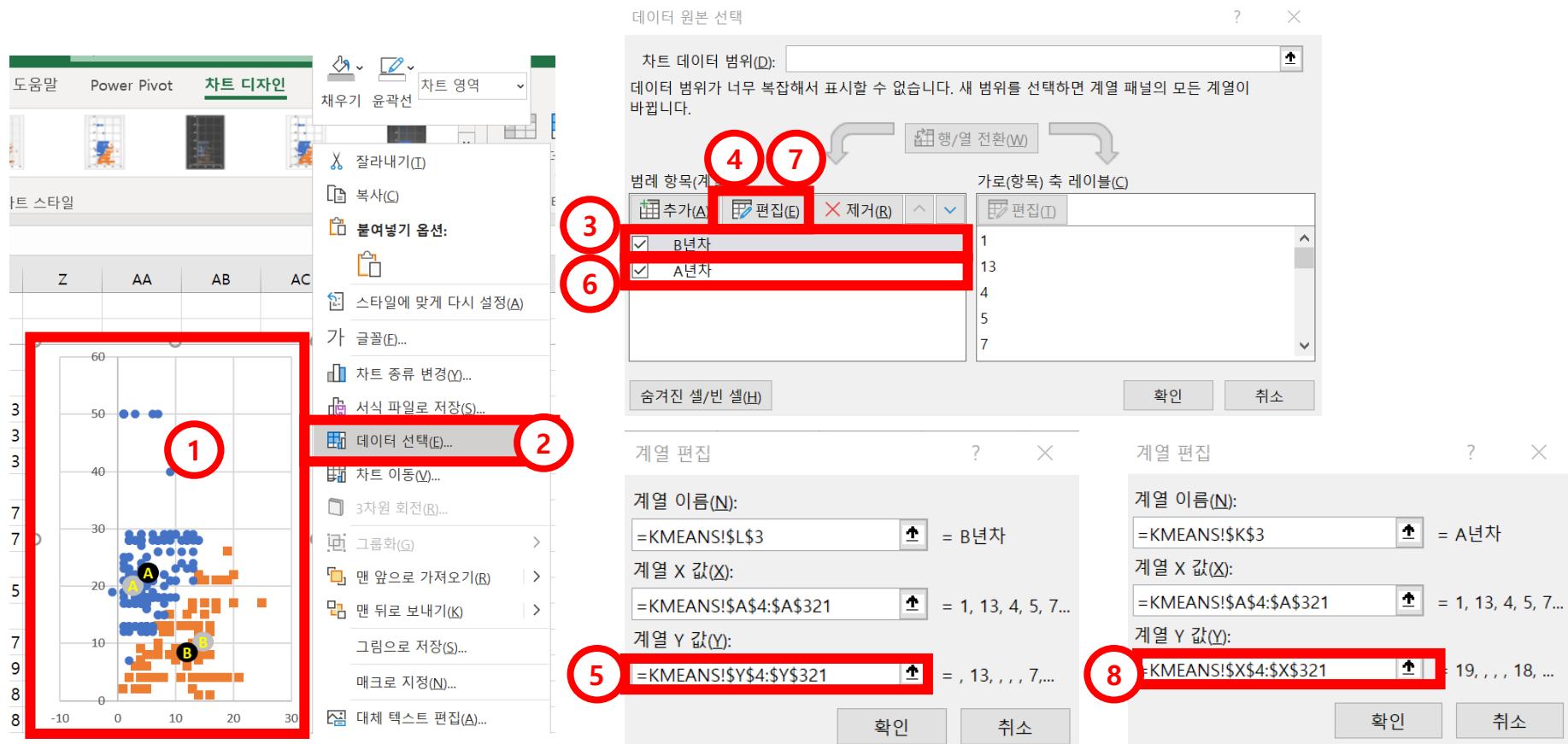
- A그룹의 경우 처음에 (2, 20)으로 추정했던 중심점 위치가 (5, 22)으로 이동했습니다.
- B그룹의 경우도 처음에 (15, 10)로 추정했던 중심점 위치가 (11,9)로 이동했습니다.
- 이제 산점도의 A그룹, B그룹 데이터 영역을 2단계값으로 수정하겠습니다.



2. 비지도학습 : K-means

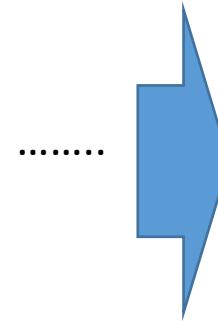
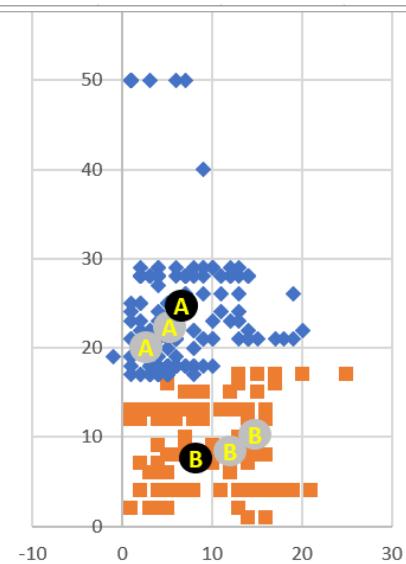
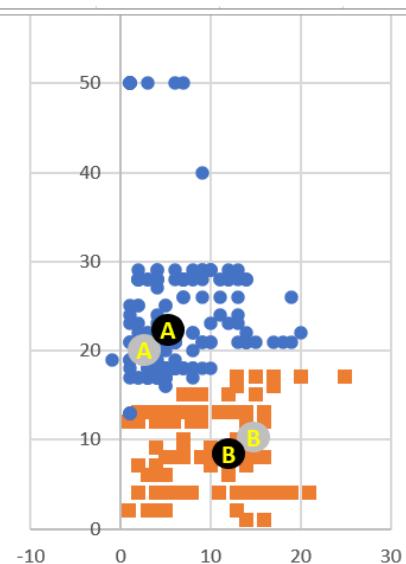
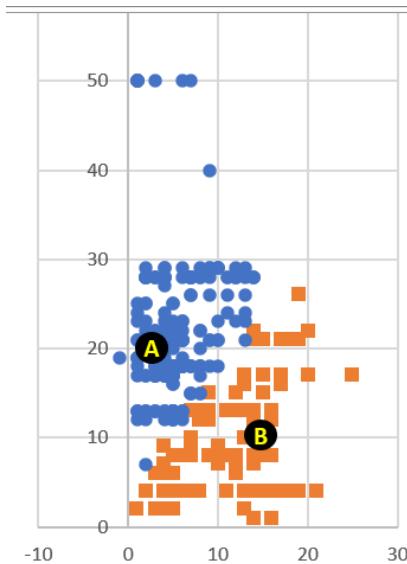
60

- 2번째 차트 선택 > 오른쪽 마우스 > 데이터 선택
- B년차 선택 > 편집 클릭 > 계열 Y값을 2단계에서 새로 그루핑한 Y열로 변경
- A년차 선택 > 편집 클릭 > 계열 Y값을 2단계에서 새로 그루핑한 X열로 변경



2. 비지도학습 : K-means

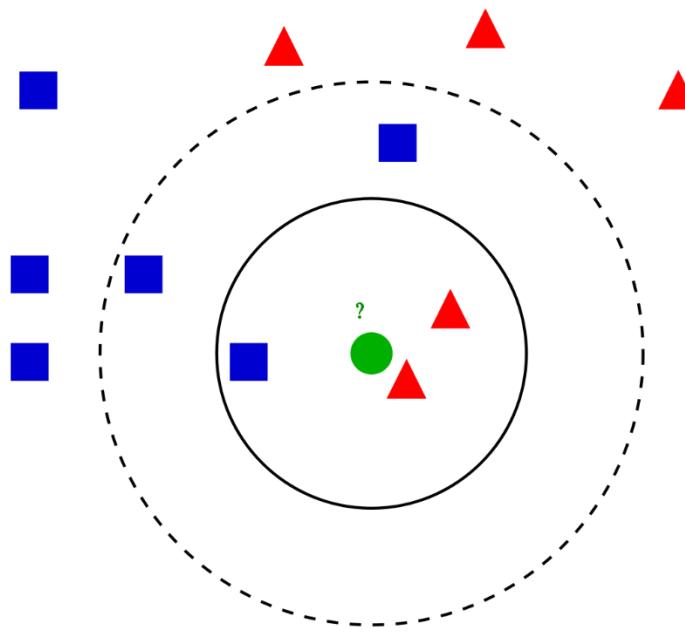
- 같은 요령으로 3단계도 작업합니다.
- K-means에서 군집을 나누는 방법은 이렇게
중심점을 구하고 -> 그루핑 -> 각 그룹의 새로운 중심점(평균)을 구하는 단계를 지정 횟수 또는 수렴이 될 때까지 반복합니다.



3. 지도학습 : KNN

62

- KNN은 원래 분류분석을 위한 알고리즘으로 주어진 조건(X값들, 독립변수)의 위치에서 가장 가까운 거리에 있는 K개의 데이터를 뽑은 다음
- 어떤 분류의 개수가 가장 많은지를 측정하여 해당값의 분류값을 결정하는 방식입니다.
- 회귀분석에 사용할 때는 K개의 가까운 데이터를 뽑은 다음 이 K개의 데이터가 가진 Y값(목표변수)의 평균을 구합니다.



* 출처 : https://ko.wikipedia.org/wiki/K-최근접_이웃_알고리즘

3. 지도학습 : KNN

63

- 부동산거래 데이터 중 거래금액, 전용면적, 층 변수로 거래금액을 예측하는 KNN 알고리즘을 계산해보겠습니다.
- (1) 먼저 'KNN' 시트를 만듭니다.
- (2) '복사본' 시트에서 거래금액, 전용면적, 층 데이터를 복사한 후 붙여넣기 합니다. 예전에 발생한 거래데이터를 기초로 새로운 전용면적과 층의 아파트가 있을 때 적정 거래가격이 얼마인지 예측합니다. 즉, 3개의 열 중 2개의 열인 전용면적과 층 값을 가지고 나머지 열 값인 거래금액을 예측하는 것입니다.
- 새로운 아파트의 (3)전용면적을 100, 층을 8이라고 해보겠습니다. K값은 인접한 몇 개의 사례를 사용할 것인지를 지정하는 값입니다. K값을 5로 입력하여 사례 중 인접한 5개 데이터를 사용하겠습니다. 기준 전용면적 100과 층 8과 각 행의 데이터 사이의 거리를 계산하기 위해 유clidean 거리를 계산합니다.
- (4) D2셀에 `=SQRT((B2-\$G\$1)^2+(C2-\$I\$1)^2)`를 입력한 후 아래로 채워줍니다. 각 행의 데이터와 기준값인 전용면적 100, 층 8과의 거리가 계산되어 채워졌습니다.

The screenshot shows an Excel spreadsheet with the following data and annotations:

	A	B	C	D	E	F	G	H	I	J
1	거래금액	전용면적	층	거리	3	전용면적	100	층	8	
2	80000	84.82	1	16.71623163	3	K				
3	209000	163.33	13	63.52707218						
4	160000	158.99	4	59.12546067						
5	96000	103	5	16.3077						
6	32000	48.54	7	51.46971537						
7	19700	16.98	4	83.11630646						
8	20000	16.98	12	83.11630646						
9	120000	84.9	7	15.13307636						

Annotations:

- Cell D2 contains the formula `=SQRT((B2-G1)^2+(C2-I1)^2)`. The formula bar is highlighted with a red box, and the number 4 is circled in red.
- Row 1 contains column headers: 거래금액, 전용면적, 층, 거리, 전용면적, 층.
- Cells G1 and H1 contain the values 100 and 8 respectively, which are circled in red.
- Cell E5 contains the value 16.3077, which is circled in red.
- Cell D5 contains the value 16.3077, which is circled in red.
- Cell D2 is highlighted with a red box.
- Cell F1 is highlighted with a red box.
- Cell G1 is highlighted with a red box.
- Cell H1 is highlighted with a red box.
- Cell I1 is highlighted with a red box.
- Cell J1 is highlighted with a red box.
- The tab bar at the bottom shows tabs for 실거래가2011, 차트, 복사본, 상관분석, 공분산과상관계수, KMEANS, and KNN. The KNN tab is highlighted with a red box and circled in red.

3. 지도학습 : KNN

- 계산후 > 상단메뉴 > 데이터 > 정렬 > 유클리디안거리 오름차순으로 정렬합니다.

3. 지도학습 : KNN

65

- 가장 가까운 거래금액 5개의 평균을 구합니다.

Screenshot of Microsoft Excel showing a dataset for KNN analysis. The 'Data' tab is selected in the ribbon. The formula bar shows A2: 85000. The data table has columns: 거래금액 (Transaction Amount), 전용면적 (Useable Area), 층 (Floor), 거리 (Distance), and 예상거래금액 (Predicted Transaction Amount). The last column is calculated using the formula =AVERAGE(A2:A6).

	A	B	C	D	E	F	G	H	I
1	거래금액	전용면적	층	거리	전용면적	100	층		
2	85000	100.77		6	2.143105224				
3	129500	97.61		7	2.590772086				
4	220000	01.991		6	2.822070339				
5	134000	95.88		5	5.0908609				
6	134500	95.88		5	5.096508609				
7	46300	94.94		7	5.157867777				
8	138000	94.51		8	5.49				

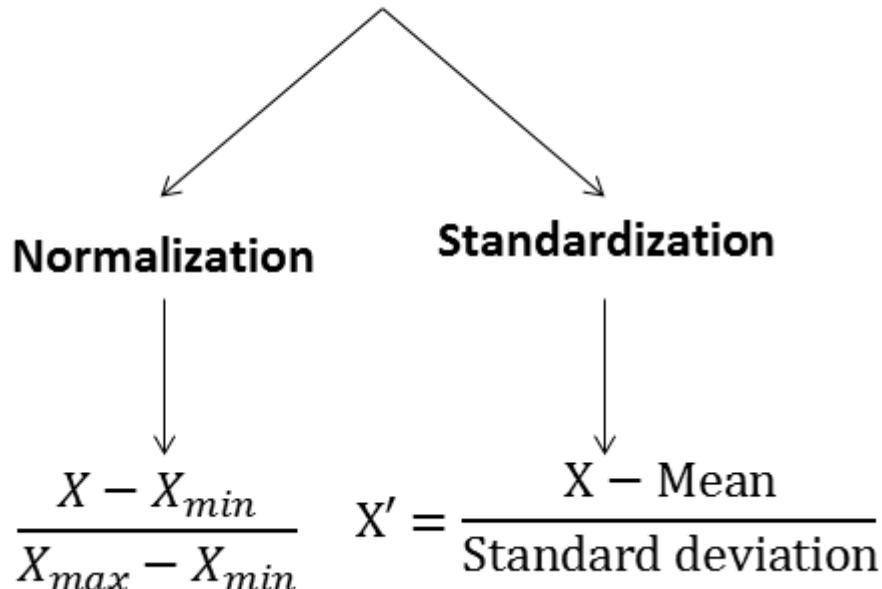
The cells A2, A6, and the formula cell H2 are highlighted with red boxes and circled with red numbers 1 and 2 respectively. The formula cell H2 contains the text '=AVERAGE(A2:A6)'.

3. 지도학습 : KNN

66

- 변수 스케일링

Feature scaling



3. 지도학습 : KNN

67

- 정규화를 진행해서 차이를 비교해봅니다. 정규화의 문제는 거리 기반의 알고리즘에 발생합니다.(예:kmeans)

	A	B	C	D	E	F	G	H	I
1	거래금액	전용면적	층	유클리디안거리	전용면적	0.387881	층	0.346	
2	138000	0.36	0.35	0.02	K		5		
3	105000	0.42	0.35	0.03					
4	129500	0.38	0.31	0.04	예상거래금액	105,360			
5	46300	0.37	0.31	0.04					

A	B	C
1 거래금액	전용면적	
2 138000	=KNN!B8-MIN(KNN!B\$2:B\$319))/(MAX(KNN!B\$2:B\$319)-MIN(KNN!B\$2:B\$319))	=KNN!C8-MIN(KNN!C\$2:C\$319))/(MAX(KNN!C\$2:C\$319)-MIN(KNN!C\$2:C\$319))
3 105000	=KNN!B11-MIN(KNN!B\$2:B\$319))/(MAX(KNN!B\$2:B\$319)-MIN(KNN!B\$2:B\$319))	=KNN!C11-MIN(KNN!C\$2:C\$319))/(MAX(KNN!C\$2:C\$319)-MIN(KNN!C\$2:C\$319))
4 129500	=KNN!B3-MIN(KNN!B\$2:B\$319))/(MAX(KNN!B\$2:B\$319)-MIN(KNN!B\$2:B\$319))	=KNN!C3-MIN(KNN!C\$2:C\$319))/(MAX(KNN!C\$2:C\$319)-MIN(KNN!C\$2:C\$319))

D	E	F	G	H	I
1 유클리디안거리	전용면적		=KNN!G1-MIN(KNN!B\$2:B\$319))/(MAX(KNN!B\$2:B\$319)-MIN(KNN!B\$2:B\$319))	층	=KNN!I1-MIN(KNN!C\$2:C\$319))/(MAX(KNN!C\$2:C\$319)-MIN(KNN!C\$2:C\$319))
2 =SQRT((B2-\$G\$1)^2+(C2-\$I\$1)^2)	K	5			
3 =SQRT((B3-\$G\$1)^2+(C3-\$I\$1)^2)					
4 =SQRT((B4-\$G\$1)^2+(C4-\$I\$1)^2)	예상거래금액	=AVERAGE(A2:A6)			
5 =SQRT((B5-\$G\$1)^2+(C5-\$I\$1)^2)					

3. 지도학습 : 회귀분석

- 회귀분석을 할 데이터를 복사하고, 시트 이름을 회귀분석으로 수정합니다.
- 데이터 > 데이터 분석 > 회귀분석 선택

1 회귀분석

2 A1:D18

3 데이터

4 데이터 분석

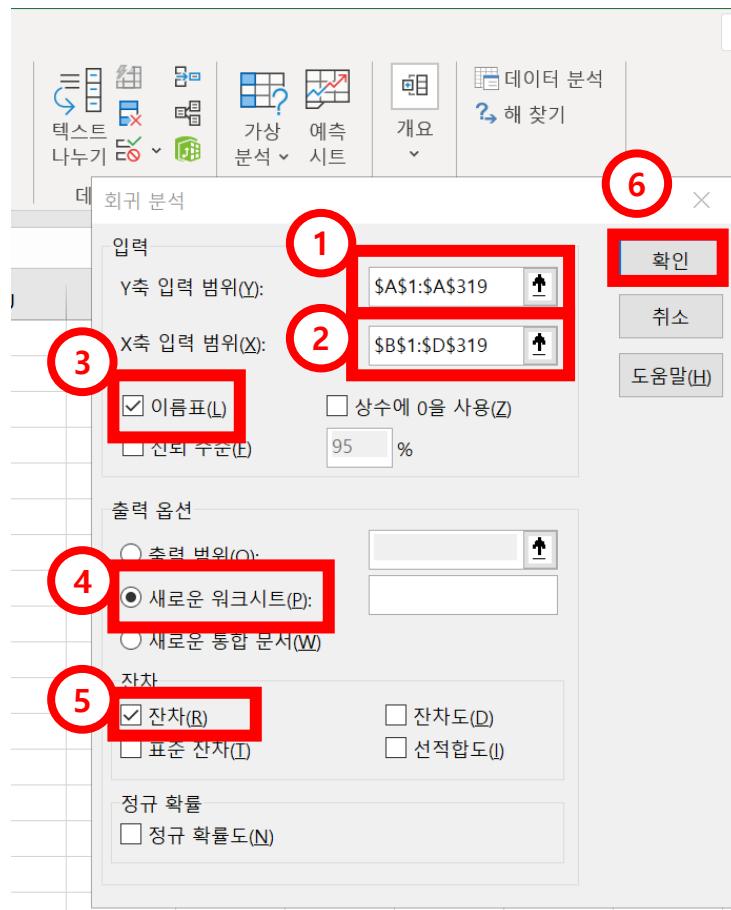
5 회귀 분석

6 확인

	거래금액	전용면적	층	년차
1	80000	84.82	1	19
2	209000	163.33	13	13
3	160000	158.99	4	13
4	96000	116.03	5	13
5	32000	48.54	7	18
6	19700	16.98	4	7
7	20000	12	7	
8	120000	7	26	
9	84500	84.67	12	15
10	67700	54.7	3	28
11	127500	111.73	17	17
12	11600	15	7	9
13	13000	17.811	5	8
14	12500	17.811	6	8
15	12650	17.811	7	8
16	197000	84.614	5	4
17	199500	84.836	14	4

3. 지도학습 : 회귀분석

- Y입력범위를 거래금액열로 지정 > X입력범위를 전용면적~년차열로 지정 > 이름표 체크 > 새로운 워크시트 > 잔차 체크한 후 > 확인



3. 지도학습 : 회귀분석

- 결정계수도 높은 편이고, 유의한 F나 각 변수의 P값도 유의한 결과가 나왔습니다.
- 회귀식은 $Y = 962.1507 * \text{전용면적} + 2058.152 * \text{층} - 1925.69 * \text{년차} + 36709.51$

요약 출력							
회귀분석 통계량							
다중 상관계수	0.804771						
결정계수	0.647656	①					
조정된 결정계수	0.64429						
표준 오차	35117.31						
관측수	318						
분산 분석							
	자유도	제곱합	제곱 평균	F 비	유의한 F		
회귀	3	7.12E+11	2.37E+11	192.3915	8.54E-71	②	
잔차	314	3.87E+11	1.23E+09				
계	317	1.1E+12					
	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0% 상위 95.0%
Y 절편	36709.51	7085.955	5.180601	3.96E-07	22767.55	50651.46	22767.55 50651.46
전용면적	962.1507	47.36664	20.31284	3.7E-59	868.9546	1055.347	868.9546 1055.347
층	2058.152	417.7159	4.927159	1.35E-06	1236.276	2880.028	1236.276 2880.028
년차	-1925.69	212.6156	-9.05715	1.46E-17	-2344.02	-1507.36	-2344.02 -1507.36

3. 지도학습 : 회귀분석

- 회귀분석 결과가 정말 회귀식으로 계산한 결과와 동일한 값을 보여주는지 실습하겠습니다.
- (1)회귀분석 결과가 표시된 시트의 이름을 '회귀분석결과'로 변경합니다.
- (2)회귀분석결과 시트 아래쪽에 있는 예측 결과값과 잔차를 복사한 후
- (3)회귀분석 시트로 이동하여 년차열 오른쪽에 (4)붙여넣기 합니다.

A	B	C	
23			
24	잔차 출력		
25			
26	관측수	측치 거래금	잔차
27	1	83789.14	-3789.14
28	2	195579.6	13420.42
29	3	172880.5	-12880.5
30	4	133604.6	-37604.6
31	5	63156.92	-31156.9
32	6	47799.6	28099.6
33	7	64264.81	-44264.8
34	8	82735.19	37264.81
35	9	113987.3	-29487.3
36	10	41594.25	26105.75
37	11	146462.4	-18962.4
38	12	48217.61	-36617.6
39	13	48731.6	-35731.6
40	14	50789.76	39289.9

1

회귀분석결과

A	B	C	D	E	F	G
1	거래금액	전용면적	층	년차	예측치	거
2	80000	84.82	1	19	83789.14	-3789.14
3	209000	163.33	13	13	195579.6	13420.42
4	160000	158.99	4	13	172880.5	-12880.5
5	96000	116.03	5	13	133604.6	-37604.6
6	32000	48.54	7	18	63156.92	-31156.9
7	19700	16.98	4	7	47799.6	-28099.6
8	20000	16.98	12	7	64264.81	-44264.8
9	120000	84.9	7	26	82735.19	37264.81
10	84500	84.67	12	15	113987.3	-29487.3
11	67700	54.7	3	28	41594.25	26105.75
12	127500	111.73	17	17	146462.4	-18962.4
13	11600	15	7	9	48217.61	-36617.6
14	13000	17.811	5	8	48731.6	-35731.6
15	12500	17.811	6	8	50789.76	-38289.8
16	12650	17.811	7	8	52847.91	-40197.9
17	197000	84.614	5	4	120708.9	76291.07
18	199500	84.836	14	4	130445.0	69541.11

3

회귀분석

3. 지도학습 : 회귀분석

- 회귀계수를 활용한 회귀식이 회귀분석 결과와 일치하는지 확인하기 위해서 회귀식에 따라 예측값과 오차를 구해봅니다.
- 회귀식을 실습하기 위해 가장 위에 행을 3개 만들어 둡니다. (1)앞서 구한 회귀식의 각 변수별 계수를 '회귀분석결과' 시트에서 가져옵니다. 잔차 오른쪽 열에 '계산금액'열과 '계산잔차'열을 만든 다음 계산금액열의 값으로 (2)G5셀에 `=\$B\$2*B5+\$C\$2*C5+\$D\$2*D5+\$E\$2`를 입력합니다. 회귀분석에서 추출한 계수와 실제값을 각각 곱하여 더해준 값을 계산합니다. (3)H5셀에는 이 계산금액값과 실제값의 차이를 구하기 위해 `=A5-G5`를 입력합니다.
- G5셀과 H5셀을 복사하여 아래 행에도 채워줍니다. 이렇게 계산한 값이 왼쪽에 있는 회귀분석결과 시트의 내용과 같은지 확인합니다.

				G		H
A	B	C	D	F	G	H
1	전용면적	층	년차	절편		
2	회귀식계산	=회귀분석결과!B18	=회귀분석결과!B19	=회귀분석결과!B20	=회귀분석결과!B17	
3						
4	거래금액	전용면적	층	년차	예측치 거래금액	잔차
5	80000	84.82	1	19	83789.144845157	-3789.14484515699
6	209000	163.33	13	13	195579.577509781	13420.4224902195
7	160000	158.99	4	13	172880.47144079	-12880.4714407896
8	96000	116.03	5	13	133604.628367577	37604.6283675768
9	32000	48.54	7	18	63156.9223976372	31156.9223976372
10	19700	16.98	4	7	47799.5952039061	-28099.5952039061
11	20000	16.98	12	7	64264.8146583253	-44264.8146583253
12	120000	84.9	7	26	82735.1905196441	37264.8094803559

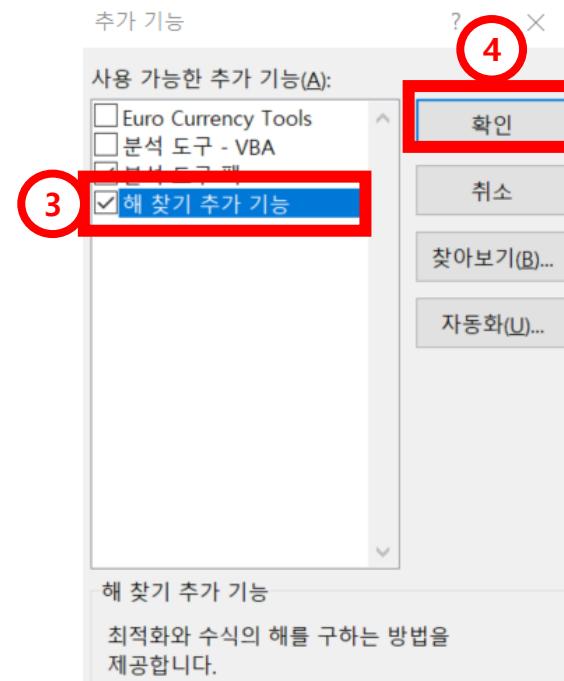
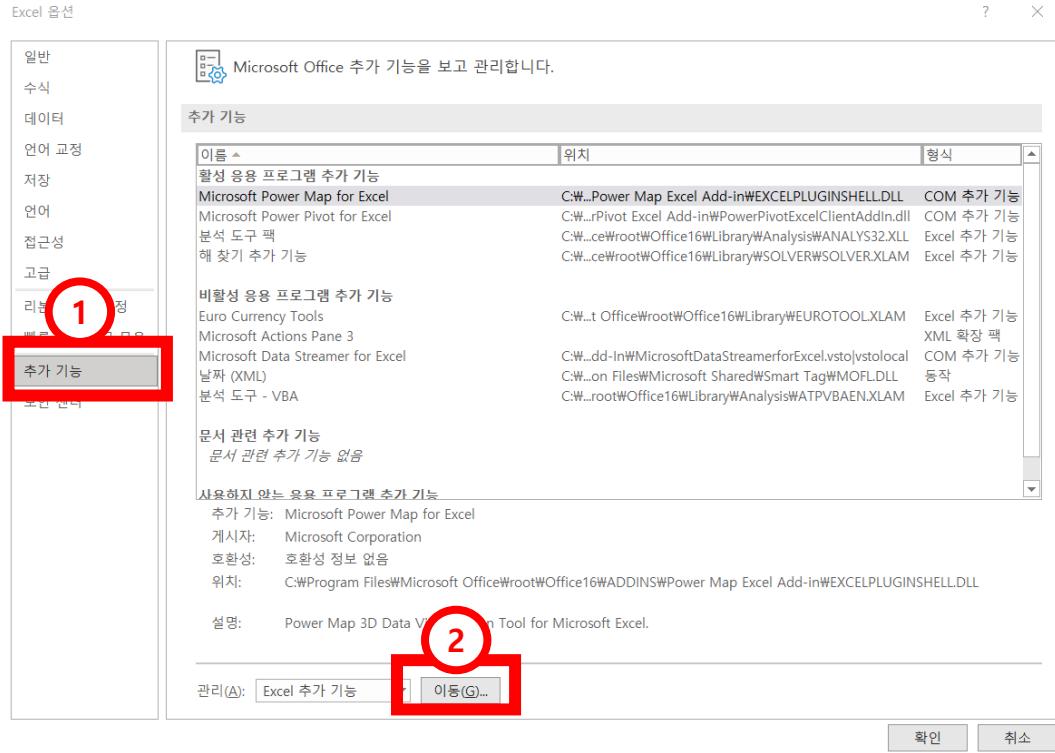
3. 지도학습 : 회귀분석

- 이번에는 회귀분석의 절차를 확인해보기 위해서 '해찾기' 기능을 활용하여 회귀계수를 계산합니다.
 - 우선 회귀계수를 모르는 상태에서 회귀계수를 찾아야 하기 때문에 동일하게 1,000으로 설정합니다.
 - 회귀식은 예측값과 실제값 사이의 잔차²을 합한 값을 최소화해야 하기 때문에, 실제로 임의로 지정한 회귀계수를 사용하여 잔차²의 합을 구해봅니다.

SUM	▼	✖	✓	fx	=B\$3*B6+\$C\$3*C6+\$D\$3*D6+\$E\$3							
A	B	C	D	E	F	G	H	I	J	K	L	
1		전용면적	층	년차	절편							
2	회귀식계산	962	2,058	-1,926	36,710							
	지정값계산	1,000	1,000	1,000	1,000							
4												
5	거래금액	전용면적	층	년차	예측치 거래금액	잔차	계산금액	계산잔차	지정금액	지정잔차	지정잔차제곱	지정잔차제곱의 합
6	80000	84.82	1	19	83789.14485	-3789.14485	83789.14485	-3789.144845	=B\$3*B6+\$C\$3*C6+\$D\$3*D6+\$E\$3	=A6-I6	=J6^2	=SUM(K6:K323)
7	209000	163.33	13	13	195579.5775	13420.42249	195579.5775	13420.42249		190330	18670	348568900
8	160000	158.99	4	13	172880.4714	-12880.47144	172880.4714	-12880.47144		176990	-16990	288660100
9	96000	116.03	5	13	133604.6284	-37604.62837	133604.6284	-37604.62837		135030	-39030	1523340900
10	32000	48.54	7	18	63156.9224	-31156.9224	63156.9224	-31156.9224		74540	-42540	1809651600
11	19700	16.98	4	7	47799.5952	-28099.5952	47799.5952	-28099.5952		28980	-9280	86118400
12	20000	16.98	12	7	64264.81466	-44264.81466	64264.81466	-44264.81466		36980	-16980	288320400
13	120000	84.9	7	26	82735.19052	37264.80948	82735.19052	37264.80948		118900	1100	1210000

3. 지도학습 : 회귀분석

- 파일 > 옵션 > 추가기능을 선택한 후 > 이동 버튼 클릭하여 추가기능창을 엽니다.
- 해 찾기 추가 기능을 체크한 후 > 확인 버튼 클릭하여 해찾기 기능을 활성화합니다.



3. 지도학습 : 회귀분석

- 데이터 > 해 찾기 실행한 후
- 해 찾기 매개변수 창에서 > 목표설정에 지정잔차^2의 합이 계산된 \$L\$6 선택 > 변경할 대상인 변수 셀 변경에 전용면적~절편까지의 지정값이 있는 \$B\$3:\$E\$3을 입력합니다.
- 지정잔차^2합이 가장 적은 값을 찾는 것이므로, 대상은 최소로 선택합니다.

The screenshot shows the Microsoft Excel interface with the following elements highlighted:

- 1**: The "Data" tab in the ribbon is selected.
- 2**: The "Solver Parameters" dialog box is open, with the "Find" button highlighted.
- 3**: The "Target cell" input field contains the formula $\$L\6 .
- 4**: The "By changing cells" input field contains the range $\$B\$3:\$E\3 .
- 5**: The "Optimization" dropdown is set to "Min".
- 6**: The "Constraint" section is expanded, showing the condition "Not equal to" and the value "0".
- 7**: The "Engine" dropdown is set to "GRG Nonlinear".
- 8**: The "Solve" button is highlighted.

The underlying data table is as follows:

	A	B	C	D	
1		전용면적	총	년차	절편
2	회귀식계산	962	2,058	-1,926	
3	지정값계산	1,000	1,000	1,000	
4	거래금액	전용면적	총	년차	예측치
5	80000	84.82	1	19	83
6	209000	163.33	13	13	19
7	160000	158.99	4	13	17
8	96000	116.03	5	13	
9	32000	48.54	7	18	
10	19700	16.98	4	7	
11	20000	16.98	12	7	64
12	120000	84.9	7	26	82
13	84500	84.67	12	15	11
14	67700	54.7	3	28	41
15	127500	111.73	17	17	14
16	11600	15	7	9	48
17	13000	17.811	5	8	
18		48/31.00555	-55/31.00555	48/31.00555	-55/31.00555
		51811	-18811	555853721	

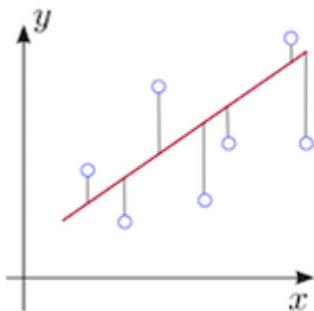
The status bar at the bottom shows tabs for "회귀분석결과", "회귀분석" (which is selected), and "회귀분석결과2".

3. 지도학습 : 회귀분석

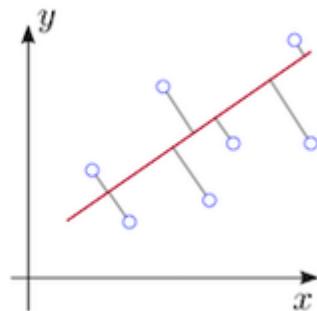
- 회귀분석을 한 결과와 거의 유사한 결과값이 나온 것을 확인하실 수 있습니다.

	A	B	C	D	E	F	G	H	I	J	K	L
1		전용면적	층	년차	절편							
2	회귀식계산	962	2,058	- 1,926	36,710							
3	지정값계산	962	2.058	- 1.926	36.710							
4												
5	거래금액	전용면적	층	년차	결과값	잔차	계산값	계산잔차	지정값	지정잔차	지정잔차^2	SUM(지정잔차^2)
6	80000	84.82	1	19	83789.14485	-3789.144845	83789.14485	-3789.144845	83789.19	-3789.19	14357944.95	3.87233E+11
7	209000	163.33	13	13	195579.5775	13420.42249	195579.5775	13420.42249	195579.4	13420.64	180113686.1	
8	160000	158.99	4	13	172880.4714	-12880.47144	172880.4714	-12880.47144	172880.3	-12880.3	165901509.7	
9	96000	116.03	5	13	133604.6284	-37604.62837	133604.6284	-37604.62837	133604.6	-37604.6	1414104093	
10	32000	48.54	7	18	63156.9224	-31156.9224	63156.9224	-31156.9224	63157.08	-31157.1	970763723.8	
11	19700	16.98	4	7	47799.5952	-28099.5952	47799.5952	-28099.5952	47799.89	-28099.9	789603742.9	
12	20000	16.98	12	7	64264.81466	-44264.81466	64264.81466	-44264.81466	64265.1	-44265.1	1959398887	
13	120000	84.9	7	26	82735.19052	37264.80948	82735.19052	37264.80948	82735.21	37264.79	1388664594	
14	84500	84.67	10	15	112007.26520	20107.26520	112007.26520	20107.26520	112007.2	20107.2	260501000.1	

OLS



TLS



보통최소제곱

ordinary least squares (OLS)

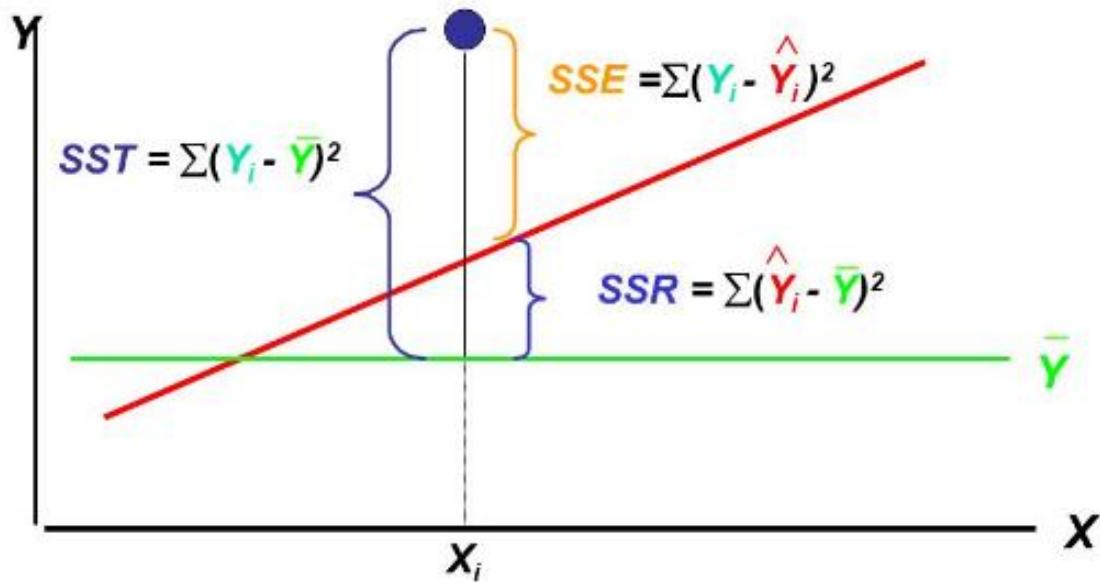
완전최소제곱

total least squares (TLS)

3. 지도학습 : 회귀분석

77

- 결정계수는 아래와 같은 수식으로 계산됩니다.
- 즉, 실제값이 평균에서 떨어진 정도와 유사한 거리만큼 예측값이 위치해야 높게 나타납니다.



$$SST = SSR + SSE$$

Total Sum of Squares

Regression Sum of Squares

Error Sum of Squares

$$\text{결정계수}(R^2) = 1 - \frac{SSE}{SST}$$

* 출처 : <https://neocarus.tistory.com/entry/선형회귀의-적합성-평가방법>

3. 지도학습 : 회귀분석

78

- 회귀분석 결과를 다시 살펴보면, 중간에 분산분석 영역이 있고, 여기에 회귀, 잔차, 계로 표시된 값이 있습니다. 이 값이 결정계수를 구하는데 사용되는 SSR, SSE, SST입니다.
1- 잔차 제곱합 나누기 계 제곱합으로도 결정계수를 구할 수 있습니다.

	A	B	C	D	E	F	G
1	요약 출력						
2							
3	회귀분석 통계량						
4	다중 상관계수	0.804771					
5	결정계수	0.647656	0.647656 = 1 - C13/C14				
6	조정된 결정계수	0.64429					
7	표준 오차	35117.31					
8	관측수	318					
9							
10	분산 분석						
11		자유도	제곱합	제곱 평균	F 비	유의한 F	
12	회귀	3	7.12E+11	2.37E+11	192.3915	8.54E-71	SSR
13	잔차	314	3.87E+11	1.23E+09			SSE
14	계	317	1.1E+12				SST

$$\begin{aligned} R_{\text{adj}}^2 &= 1 - (1 - R^2) \times \frac{n - 1}{n - k - 1} \\ &= R^2 \times \frac{n - 1}{n - k - 1} - \frac{k}{n - k - 1} \\ &= R^2 - \frac{k}{n - k - 1} (1 - R^2) \\ &\leq R^2 \end{aligned}$$

조정된 결정계수

3. 지도학습 : 회귀분석

- 데이터에 기반해서 SSR과 SST를 구해보겠습니다.

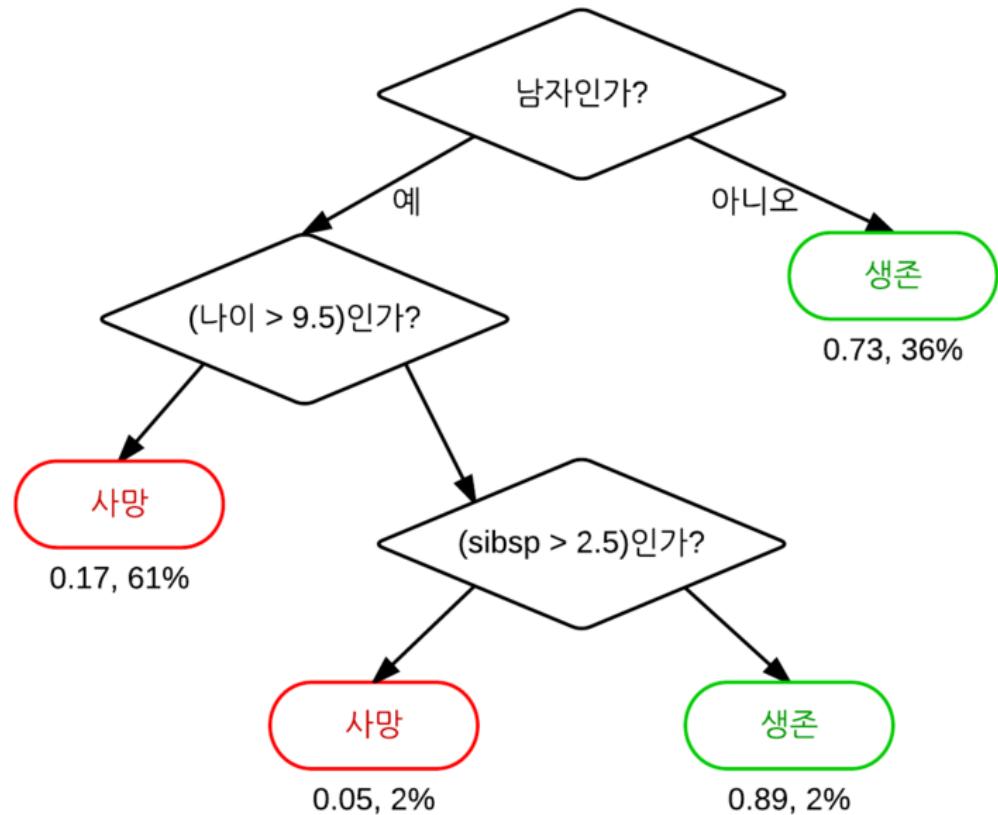
A	B	C	D	E	F	G	H	I	J
1 거래금액	전용면적	층	년차	예측치 거래금액	(거래금액-평균)제곱	(거래금액-회귀예측값)제곱	거래금액평균	4	
2 80000	84.82	1	19	83789.14485	= (A2-\$I\$2)^2		97,886	= AVERAGE(A2:A319)	
3 209000	163.33	13	13	195579.5775	12,346,214,774	180,107,740	SST	1,098,699,266,961	= SUM(F2:F319) 7
4 160000	158.99	4	13	172880.4714	3,858,089,617	165,906,545	SSE	387,218,438,934	= SUM(G2:G319) 8
5 96000	116.03	5	13	133604.6284	3,558,799	1,414,108,075	결정계수	0.64756649	= 1-I7/I5
6 32000	48.54	7	18	63156.9224	4,341,027,982	970,753,813			
7 19700	16.98	4	7	47799.5952	6,113,125,340	789,587,251			
8 20000	16.98	12	7	64264.81466	6,066,303,453	1,959,373,817			
9 120000	84.9	3	26	82735.19052	4,007,856	1,666,026			
10 84500	84.67	12	15	113987.2653	179,197,793	869,498,812			
11 67700	54.7	3	28	41594.24551	911,223,453	681,510,418			
12 127500	111.73	17	17	146462.4431	876,960,686	359,574,249			
13 11600	15	7	9	48217.61091	7,445,356,283	1,340,849,429			
14 13000	17.811	5	8	48731.60333	7,205,714,145	1,276,747,476			
15 12500	17.811	6	8	50789.75576	7,290,850,623	1,466,105,396			
16 12650	17.811	7	8	52847.90819	7,265,257,180	1,615,871,823			
17 197000	84.614	5	4	120708.925	9,823,490,246	5,820,328,119			
18 199500	84.836	14	4	139445.8944	1,035,307,856	3,606,101			

2

1

3. 지도학습 : 의사결정나무

80



* 출처 : https://ko.wikipedia.org/wiki/결정_트리_학습법

3. 지도학습 : 의사결정나무

81

- 어떻게 나눌 것인가?

이분법 속성일 때의 지니 계수 (Gini Index of Binary Attributes)



* 출처 : <https://lucy-the-marketer.kr/ko/growth/decision-tree-and-impurity/>

3. 지도학습 : 의사결정나무

82

- 거래금액이 중위값 이상이면 A그룹(고가아파트), 중위값 미만이면 B그룹(저가아파트)으로 분류하는 경우를 가지고 실습하겠습니다.
- (1)'의사결정나무'시트를 만듭니다. (2)'복사본'시트에서 거래금액, 전용면적, 층을 복사하여 (3)붙여넣습니다. (4)F2셀에 `=MEDIAN(A2:A319)`라고 수식을 입력하여 기준이 될 거래금액의 중위값을 구합니다. 중위값이 87,250(단위가 만원이므로 8억7천250만원)입니다. 이 중위값을 기준으로 A그룹과 B그룹으로 나눠보겠습니다. '금액구분'이라는 열을 D열에 만들고, (5)D2셀에 `=IF(A2>=\$F\$2,"A","B")`을 입력하고 아래 행에도 채워줍니다. 각 행의 거래금액에 따라 A그룹과 B그룹으로 나눈것을 확인하실 수 있습니다.

	A	B	C	D	E	F	G
1	거래금액	전용면적	층	금액구분		거래금액중위값	
2	85000	100.77	6	B	5	87,250	=MEDIAN(A2:A319)
3	129500	97.61	7	A			
4	220000	101.991	6	A			
5	134000	95.88	5	A			
6	134500	95.88	5	A			
7	46300	94.94	7	B			
8	138000	94.51	8	A			
9	140000	94.51	5	A			
10	134000	94.51	4	A			

3. 지도학습 : 의사결정나무

83

- (1) 우선 전체 데이터의 개수를 구합니다. F5셀에 `=COUNT(A2:A319)`를 입력하여 구합니다.
- (2)'금액구분'열에 있는 A의 개수가 총 몇 개인지 구합니다. G5셀에 `=COUNTIF(\$D\$2:\$D\$319, "A")`를 입력하여 구합니다. countif함수는 대상 범위 중에 조건에 해당하는 값이 몇 개나 있는지 세줍니다.
- (3)B의 개수도 구합니다. H5셀에 `=COUNTIF(\$D\$2:\$D\$319, "B")`를 입력하여 구합니다.
- 마지막으로 계산 공식에 의해 지니불순도를 계산합니다. 아래 공식에 따라 I6셀에 `=1-(G5/F5)^2-(H5/F5)^2`를 입력하여 구합니다.

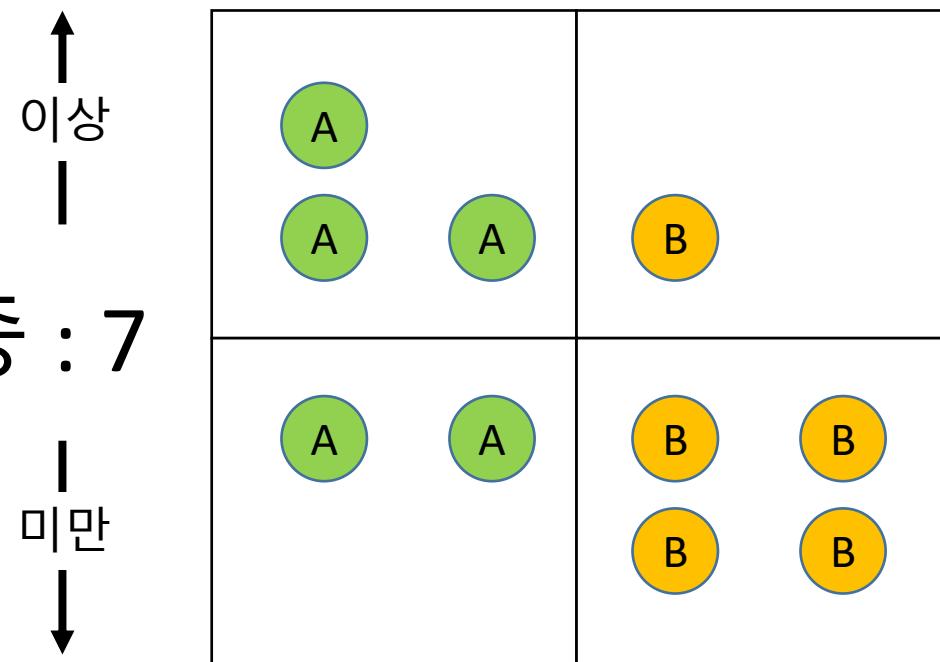
$$\begin{aligned}\text{지니불순도} &= 1 - \left(\frac{\text{A개수}}{\text{전체개수}} \right)^2 - \left(\frac{\text{B개수}}{\text{전체개수}} \right)^2 \\ &= 1 - \left(\frac{159}{318} \right)^2 - \left(\frac{159}{318} \right)^2 = 1 - \left(\frac{1}{2} \right)^2 - \left(\frac{1}{2} \right)^2 = 1 - 0.25 - 0.25 \\ &= 0.5\end{aligned}$$

	A	B	C	D	E	F	G	H	I
1	거래금액	전용면적	층	금액구분	거래금액중위값				
2	85000	100.77	6	B		87,250			
3	129500	97.61	7	A	①	②	③	④	
4	220000	101.991	6	A	전체개수	A개수	B개수	지니불순도	
5	134000	95.88	5	A	318		159	159	0.50
6	134500	95.88	5	A	=COUNT(A2:A319)	=COUNTIF(\$D\$2:\$D\$319, "A")	=COUNTIF(\$D\$2:\$D\$319, "B")	=1-(G5/F5)^2-(H5/F5)^2	
7	46300	94.94	7	B					
8	138000	94.51	8	A					
9	140000	94.51	5	A					
10	134000	94.51	4	A					

3. 지도학습 : 의사결정나무

84

- 층을 기준으로 나눌 경우 기준값을 층의 중앙값인 7로 잡고 지니불순도를 계산해보겠습니다.
- 예를 들어, 아래 그림처럼 7층을 기준으로 나누면 7층 이상으로 하나를 나눌 수 있고, 7층 미만으로 하나를 나눌 수 있습니다. 가장 이상적인 경우는 7층 이상에 A그룹이나 B그룹 1가지 종류가 다 몰려있고, 7층 미만에 다른 그룹이 몰려있는 경우입니다. 층을 기준으로 나누었지만 거래금액을 2가지 그룹으로 정확하게 나눌 수 있기 때문입니다. 반대로 가장 지니불순도가 높은 상태는 A그룹과 B그룹의 개수가 비슷하여 7층 이상을 A그룹이라고 얘기할 수도 없고, B그룹이라고 얘기할 수도 없는 경우입니다.
- 아래 그림처럼 층을 기준으로 나누게 되면 7층 이상인 경우 A그룹과 B그룹이 어느 정도 섞여있게 됩니다. 7층을 기준으로 7층 이상인 경우 A그룹과 B그룹이 얼마나 섞였는지를 공식에 의해 계산합니다. 같은 방법으로 7층 미만인 경우도 계산한 후 가중평균을 구합니다. 어느 한 쪽으로 비율이 쓸리는 현상을 가중평균으로 보완하는 것입니다.



A그룹	B그룹	가중평균
$\left(1 - \left(\frac{3}{4}\right)^2\right)$	$\left(1 - \left(\frac{1}{4}\right)^2\right)$	$\times \frac{4}{10}$ 7층이상수 전체개수
$\left(1 - \left(\frac{2}{6}\right)^2\right)$	$\left(1 - \left(\frac{4}{6}\right)^2\right)$	$\times \frac{6}{10}$ 7층미만수 전체개수

$$\left(1 - \left(\frac{3}{4}\right)^2\right) + \left(1 - \left(\frac{1}{4}\right)^2\right) \times \frac{4}{10}$$

$$\left(1 - \left(\frac{2}{6}\right)^2\right) + \left(1 - \left(\frac{4}{6}\right)^2\right) \times \frac{6}{10}$$

3. 지도학습 : 의사결정나무

- 전용면적, 층에 대한 지니계수를 구해봅시다.

	A	B	C	D	E	F	G	H	I	J	K	L
1	거래금액	전용면적	층	금액구분	거래금액중위값							
2	85000	100.77	6	B		87,250						
3	129500	97.61	7	A								
4	220000	101.991	6	A	전체개수		A개수	B개수	지니불순도			
5	134000	95.88	5	A		318	159	159	0.50			
6	134500	95.88	5	A								
7	46300	94.94	7	B	[층]							
8	138000	94.51	8	A	1	구분값(중위값)	7					
9	140000	94.51	5	A								
10	134000	94.51	4	A			A개수	B개수	지니불순도	3	정보획득	
11	105000	106.81	8	A	구분값이상	2	88	7	0.49	0.50	0.49	0.01
12	89700	106.62	11	A	구분값미만	67	81	0.49	0.49			4

	F	G	H	I	J	K	L	
7	1							
8	구분값	=MEDIAN(C2:C319)						
9		2						
10	A개수							
11	구분값이상	=SUMPRODUCT((\$C\$2:\$C\$319>=\$G\$8)*(\$D\$2:\$D\$319="A")*1)	=SUMPRODUCT((\$C\$2:\$C\$319>=\$G\$8)*(\$D\$2:\$D\$319="B")*1)	=1-(G11/(G11+H11))^2-(H11/(G11+H11))^2	=SUM(G11:H11)/\$F\$5	=I11*J11+I12*K12	정보획득	
12	구분값미만	=SUMPRODUCT((\$C\$2:\$C\$319<\$G\$8)*(\$D\$2:\$D\$319="A")*1)	=SUMPRODUCT((\$C\$2:\$C\$319<\$G\$8)*(\$D\$2:\$D\$319="B")*1)	=1-(G12/(G12+H12))^2-(H12/(G12+H12))^2	=SUM(G12:H12)/\$F\$5		4	

3. 지도학습 : 의사결정나무

- 사용한 수식은 아래 그림과 같습니다.

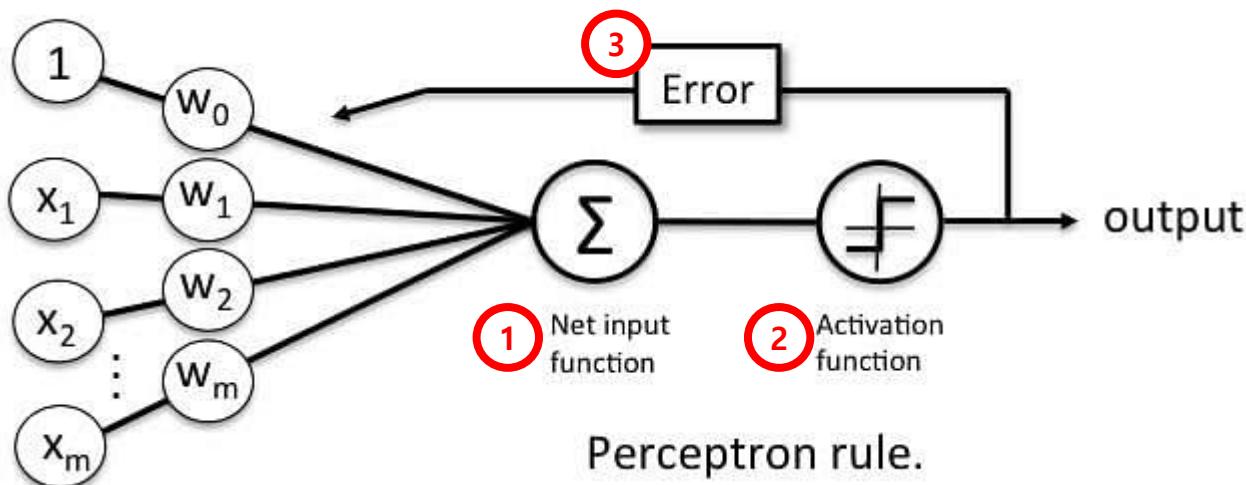
	A	B	C	D	E	F	G	H	I	J	K	L
1	거래금액	전용면적	층	금액구분		거래금액중위값						
2	85000	100.77	6	B		87,250						
3	129500	97.61	7	A								
4	220000	101.991	6	A	전체개수		A개수	B개수	지니불순도			
5	134000	95.88	5	A		318	159	159	0.50			
6	134500	95.88	5	A								
7	46300	94.94	7	B	[층]							
8	138000	94.51	8	A	구분값		7					
9	140000	94.51	5	A								
10	134000	94.51	4	A			A개수	B개수	지니불순도		정보획득	
11	105000	106.81	8	A	구분값이상		88	71	0.49	0.50	0.49	0.01
12	89700	106.62	11	A	구분값미만		67	88	0.49	0.49		
13	98500	106.62	12	A								
14	91000	107.94	7	A	【전용면적】							
15	78000	91.53	9	B	1 구분값	84						
16	144000	108.07	5	A								
17	144000	108.07	5	A			A개수	B개수	지니불순도	0.26	0.26	정보획득
18	44000	91.87	5	B	구분값이상	2	131	24	0.26	0.49	0.26	0.24
19	74000	108.18	3	B	구분값미만	3	24	135	0.26	0.50	4	

	F	G	H	I	J	K	L
14	【전용면적】	1					
15	구분값	=MEDIAN(B2:B319)					
16		2					
17	A개수		B개수				
18	구분값이상	=SUMPRODUCT((\$B\$2:\$B\$319>=\$G\$15)*(\$D\$2:\$D\$319="A")*1)	=SUMPRODUCT((\$B\$2:\$B\$319>=\$G\$15)*(\$D\$2:\$D\$319="B")*1)	=1-(G18/(G18+H18))^2-(H18/(G18+H18))^2	=SUM(G18:H18)/F\$5	=I18*J18+I19*K19	=L12-K18
19	구분값미만	=SUMPRODUCT((\$B\$2:\$B\$319<\$G\$15)*(\$D\$2:\$D\$319="A")*1)	=SUMPRODUCT((\$B\$2:\$B\$319<\$G\$15)*(\$D\$2:\$D\$319="B")*1)	=1-(G19/(G19+H19))^2-(H19/(G19+H19))^2	=SUM(G19:H19)/F\$5		

4. 딥러닝 : 퍼셉트론

87

- 딥러닝의 가장 기초적인 형태인 퍼셉트론 알고리즘의 계산 방식을 실습합니다.
- (1)퍼셉트론은 각 변수에 가중치를 두어 곱한 다음 이 값을 전부 합칩니다.(Net input function)
- (2)합쳐진 값을 기준점을 두고 분류합니다. (Activation function)
- (3)이 결과값이 실제값과 일치하는지를 판단하여 최초 설정한 가중치를 보정합니다.(Error)



* 출처 : <https://www.simplilearn.com/tutorials/deep-learning-tutorial/perceptron>

4. 딥러닝 : 퍼셉트론

88

- (1) 퍼셉트론 실습을 진행할 시트를 하나 만듭니다.
- (2) 복사본 시트에서 전용면적, 층, 거래금액 열을 복사하여 B4셀을 기준으로 붙여넣습니다. 열머리글에 나중에 사용할 변수명(X1, X2, Y)를 함께 입력합니다. 전용면적을 전용면적(X1), 층을 층(X2), 거래금액을 거래금액(Y)로 수정합니다.
- (3) X0에 해당하는 절편(머신러닝, 딥러닝에서는 BIAS로 표현)을 A열에 만듭니다. A4셀에는 열머리글인 '절편(X0)'를, A5셀부터는 1을 입력하여 마지막까지 채워줍니다.

	SUM		X	✓	f(x)	=IF(D5>=\$C\$1,1,0)	
A	B	C	D	E	F	G	H
1	거래금액중위값	(4)	87,250	=MEDIAN(D5:D322)			
2							
3							
4	절편(X0)	전용면적(X1)	층(X2)	거래금액(Y)	금액구분		
5	1	100.77	6	85000	=IF(D5>=\$C\$1,1,0)		
6	1	97.61	7	129500	1		
7	1	101.991	6	220000	1		
8	1	95.88	5	134000	1		
9	1	95.88	2	5	134500	(5)	
10	1	94.94	7	46300	0		
11	1	94.51	8	138000	1		
12	1	94.51	5	140000	1		
13	1	94.51	4	134000	1		
14	1	100.01	9	105000	1		

- (4) 퍼셉트론으로 고가아파트인지 저가아파트인지를 분류할 것이므로 이 분류를 위한 기준인 중위값을 C1셀에 계산합니다. 수식에 `=MEDIAN(D5:D322)`을 입력합니다.
- (5) E열에는 거래금액의 중위값인 C1셀값 이상이면 1(고가아파트), 미만이면 0(저가아파트)로 분류하겠습니다. E5셀에 `=IF(D5>=\$C\$1,1,0)`를 입력하고 마지막까지 채워줍니다.

4. 딥러닝 : 퍼셉트론

- (1) 3개의 가중치(W_0 , W_1 , W_2) 값을 임의로 1로 지정합니다. 퍼셉트론이 학습을 진행하면서, 이 가중치를 변화시키는 것이기 때문에 최초값에 큰 의미를 두지 않아도 됩니다.
 - (2) 1개의 절편(X_0)과 2개의 변수(X_1 , X_2)를 3개의 가중치와 각각 곱합니다. I5셀에는 절편에 대한 첫 번째 가중치인 W_0 와 절편값인 X_0 를 곱해줍니다. 수식에 `=A5*F5`를 입력합니다. J5셀에는 W_1 과 전용면적에 해당하는 X_1 을 곱해줍니다. 수식에 `=B5*G5`를 입력합니다. K5셀에는 W_2 와 층에 해당하는 X_2 를 곱해줍니다. 수식에 `=C5*H5`를 입력합니다.
 - (3) 이제 이 3가지 값을 합해줍니다. L5셀에 `=I5+J5+K5`를 입력합니다.
 - (4) 합해진 결과값이 0이상이면 1, 0미만이면 0을 얻는 조건처리를 하겠습니다. 퍼셉트론의 구조 중 Activation Function에 해당하는 부분입니다. M5셀에 `=IF(L5>=0,1,0)`를 입력합니다.
 - (5) 실제값(거래금액(Y))과 계산값을 비교하여 오차가 있는지 계산합니다. 만약 계산값이 실제값보다 크다면 값을 줄여주어야 오차가 줄어들 것이고, 계산값이 실제값보다 적다면 값을 키워주어야 오차가 줄어들 것입니다. N5셀에 `=E5-M5`를 입력하여 오차를 구합니다.

4. 딥러닝 : 퍼셉트론

- 오차가 많이 발생했네요. 이 오차에 따라 가중치를 수정해 줍니다. 계산값이 실제값보다 크면 기존 가중치에 음수를 더해주어 가중치를 줄여주고, 계산값이 실제값보다 작으면 기존 가중치에 양수를 더해주어 가중치를 키워주겠습니다.
 - 이를 위해 기존 가중치에 얼마를 더해줄지를 결정해야 합니다. 이 값은 오차(오차의 부호) * X값 * 학습률으로 계산합니다. 오차를 곱해서 가중치를 더할지 빼줄지를 결정합니다. X값의 크기가 다 제각각이기 때문에 이를 반영해주기 위해 X값을 곱해주고 마지막으로 얼마나 큰 폭으로 가중치를 수정할지를 결정하는 학습률을 곱해줍니다.
 - (1) C2셀에 학습률을 0.01로 지정하여 입력합니다. 학습률에 따라 퍼셉트론이 가중치를 조정하는 폭이 달라집니다. 1보다 적은 수로 지정하면 됩니다. 이 실습에서는 0.01로 지정하겠습니다. 보통은 0.01이하의 숫자로 지정합니다.
 - (2) 오차(오차의 부호) * X값 * 학습률을 계산합니다. 오차는 N5셀, X값은 A5, B5, C5, 학습률은 C2셀에 있습니다. O5에는 `=A5*\$N5*\$C\$2`, P5에는 `=B5*\$N5*\$C\$2`, Q5에는 `=C5*\$N5*\$C\$2` 수식을 입력하여 증감시킬 가중치를 구합니다.

4. 딥러닝 : 퍼셉트론

- (1) F6, G6, H6 셀에 각각 `=F5+O5`, `=G5+P5`, `=H5+Q5`를 입력합니다. 첫 번째 행의 오류를 반영한 수정된 가중치가 만들어졌습니다.
- (2) I열부터 Q열까지의 수식은 첫 번째 행과 동일하므로 복사한 후 붙여넣습니다.
- 그 다음 2번째 행(F6:Q6)을 아래 빈 영역에 채워줍니다.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1 거래금액증위값	87,250															
2 학습률	0.01															
3																
4 절편(X0)	전용면적(X1)	층(X2)	거래금액(Y)	금액구분	W0	W1	W2	W0*X0	W1*X1	W2*X2	i+j+k	예측값	오차	W0증감	W1증감	W2증감
5 1 100.77	6 85000	0	1.00	1.00	1.00	1.00	1.00	100.8	6.0	107.8	1 - 1	- 0.0100	- 1.0077	- 0.0600		
6 1 97.61	7 129500	1	0.99	- 0.01	0.94	0.99	- 0.8	6.6	6.8	1	-	-	-	-	-	-
7 1 101.991	6 220000	1	=F5+O5	=G5+P5	=H5+Q5											
8 1 95.88	5 134000	1														

1

2

4. 딥러닝 : 퍼셉트론

- 이렇게 전체 데이터에 대한 학습이 1회 끝난 것을 epoch라고 표현합니다.
- (1)마지막 행까지 계산을 한 마지막 가중치가 얼마인지 밑에까지 스크롤하지 않아도 확인할 수 있도록 F2, G2, H2셀에 각각 `=F322`, `=G322`, `=H322` 값을 가져옵니다.
- (2)성능이 얼마나 좋은지 알아보기 위해 정답률을 구하겠습니다. 우선 정답수를 구합니다. 정답수는 오차열에 0이 몇 개인지 세어서 구합니다. N2셀에 `=COUNTIF(\$N\$5:\$N\$322,"0")`를 입력하여 구합니다. COUNTIF함수는 대상범위 내에 조건에 맞는 셀이 몇 개인지 세어주는 함수입니다. 오답수는 오차열에 있는 -1과 1이 몇 개인지 세어서 합하면 됩니다. P2셀에 `=COUNTIF(\$N\$5:\$N\$322,"-1")+COUNTIF(\$N\$5:\$N\$322,"1")`를 입력하여 구합니다. 전체 데이터수가 318개인데 정답수와 오답수를 합하니 318로 잘 계산된 것을 확인할 수 있습니다. 정답수 / (정답수+오답수)를 계산하여 정답률을 구하겠습니다. N1셀에 `=N2/(N2+P2)`를 입력합니다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	거래금액중위값	87,250											정답률	67.61%				
2	학습률	0.01			마지막가중치	0.73	0.95	- 0.11					정답	215	오답	103		
3																		
4	절편(X0)	전용면적(X1)	층(X2)	거래금액(Y)	금액구분	W0	W1	W2	W0*X0	W1*X1	W2*X2	i+j+k	예측값	오차	W0증감	W1증감	W2증감	
5	1	100.77	6	85000	0	1.00	1.00	1.00	1.00	100.8	6.0	107.8	1	-	1	- 0.0100	- 1.0077	- 0.0600
6	1	97.61	7	129500	1	0.99	- 0.01	0.94	0.99	- 0.8	6.6	6.8	1	-	-	-	-	-
7	1	101.991	6	220000	1	0.99	- 0.01	0.94	0.99	- 0.8	5.6	5.8	1	-	-	-	-	-
8	1	95.88	5	134000	1	0.99	- 0.01	0.94	0.99	- 0.7	4.7	5.0	1	-	-	-	-	-
9	1	95.88	5	134500	1	0.99	- 0.01	0.94	0.99	- 0.7	4.7	5.0	1	-	-	-	-	-
10	1	94.94	7	46300	0	0.99	- 0.01	0.94	0.99	- 0.7	6.6	6.8	1	-	1	- 0.0100	- 0.9494	- 0.0700
11	1	94.51	8	138000	1	0.98	- 0.96	0.87	0.98	- 90.5	7.0	- 82.5	0	1	0.0100	0.9451	0.0800	

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
319	1	213.07	2	125000	1	0.73	0.95	- 0.11	0.73	202.5	- 0.2	203.0	1	-	-	-	-
320	1	215.146	9	175000	1	0.73	0.95	- 0.11	0.73	204.5	- 1.0	204.2	1	-	-	-	-
321	1	236.07	5	123000	1	0.73	0.95	- 0.11	0.73	224.3	- 0.6	224.5	1	-	-	-	-
322	1	238.858	5	210000	1	0.73	0.95	- 0.11	0.73	227.0	- 0.6	227.2	1	-	-	-	-

4. 딥러닝 : 퍼셉트론

- (1) 1 epoch가 끝나서 계산된 가중치를 복사하여 (2) 영역에 값만 붙여넣기(복사 > 오른쪽마우스 클릭 > 선택하여 붙여넣기 > '값' 선택) 합니다. (3) 최초 가중치 값이 바뀌게 되면 다시 계산이 되고(2번째 epoch) (4) 새로운 가중치가 계산됩니다. (5) 정답률이 약간 상승했습니다. 정답률은 이렇게 상승할 때도 있지만, 감소할 때도 있습니다. epoch를 계속하면 조금씩 상승합니다. 이렇게 계속해서 여러 epoch를 계산한 후 최종 결과를 얻게 됩니다.

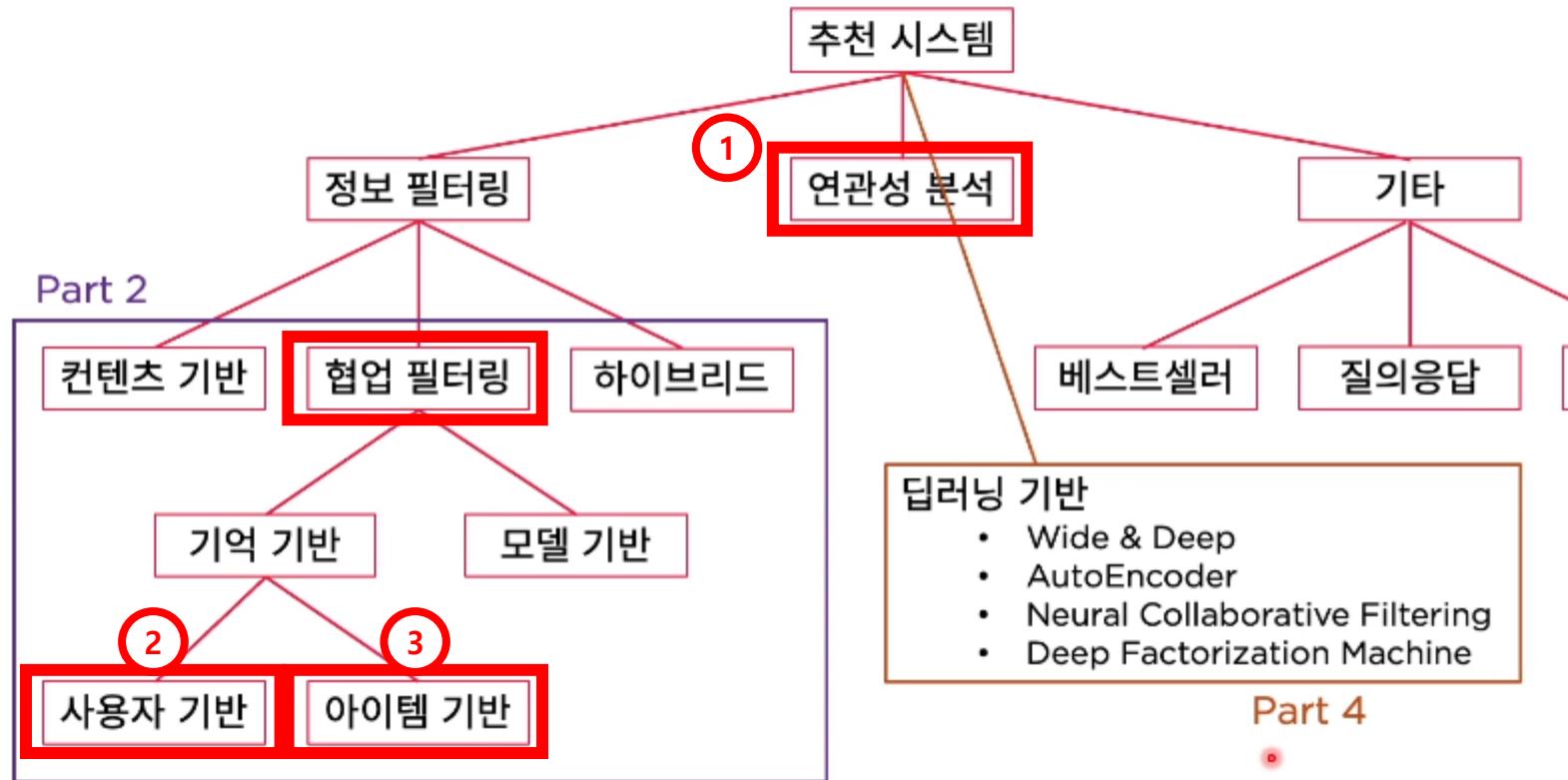
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q		
1	거래금액	중위값	87,250										정답률	67.61%	5				
2	학습률		0.01		마지막가중치	0.73	0.95	-	0.11	1			정답	215	오답	103			
3																			
4	절편(X0)	전용면적(X1)	층(X2)	거래금액(Y)	금액구분	W0	W1	W2	W0*X0	W1*X1	W2*X2	i+j+k	예측값	오차	W0증감	W1증감	W2증감		
5	1	100.77	6	85000	0	1.00	1.00	1.00	2	0	100.8	6.0	107.8	1	-	1	-0.0100	-1.0077	-0.0600
6	1	97.61	7	129500	1	0.99	-	0.01	0.94	0.99	-	0.8	6.6	6.8	1	-	-	-	
7	1	101.991	6	220000	1	0.99	-	0.01	0.94	0.99	-	0.8	5.6	5.8	1	-	-	-	

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q					
1	거래금액	중위값	87,250										정답률	67.92%	5							
2	학습률		0.01		마지막가중치	0.47	1.11	-	0.24	4			정답	216	오답	102						
3																						
4	절편(X0)	전용면적(X1)	층(X2)	거래금액(Y)	금액구분	W0	W1	W2	W0*X0	W1*X1	W2*X2	i+j+k	예측값	오차	W0증감	W1증감	W2증감					
5	1	100.77	6	85000	0	0.73	0.95	-	0.11	3	95.8	-	0.7	95.8	1	-	1	-0.0100	-1.0077	-0.0600		
6	1	97.61	7	129500	1	0.72	-	0.06	-	0.17	0.72	-	5.6	-	1.2	-	6.1	0	1	0.0100	0.9761	0.0700
7	1	101.991	6	220000	1	0.73	0.92	-	0.10	0.73	93.7	-	0.6	93.8	1	-	-	-	-	-	-	
8	1	95.88	5	134000	1	0.73	0.92	-	0.10	0.73	88.1	-	0.5	88.3	1	-	-	-	-	-	-	

5. 기타 : 추천분석

94

- 추천분석 종류



https://github.com/jangsoohoon/recommend_system/wiki/추천-알고리즘-종류

5. 기타 : 추천분석

95

- 연관분석

	A	B	C	D	E	F
1	구분	마녀2	범죄도시2	외계인1	탑건매버릭	토르러브앤썬더
2	Hyun		6	10	9	10
3	이브				7	10
4	연참		2	1		3
5	니모			8	8	9
6	경표		6		6	8
7						7
8	구분	마녀2	범죄도시2	외계인1	탑건매버릭	토르러브앤썬더
9	Hyun		1	1	1	1
10	이브				1	1
11	연참		1	1		1
12	니모			1	1	1
13	경표		1		1	1

	A	B	C	D	E	F
8	구분	마녀2	범죄도시2	외계인1	탑건매버릭	토르러브앤썬더
9	Hyun	=IF(B2="", "", IF(B2>0, 1, 0))	=IF(C2="", "", IF(C2>0, 1, 0))	=IF(D2="", "", IF(D2>0, 1, 0))	=IF(E2="", "", IF(E2>0, 1, 0))	=IF(F2="", "", IF(F2>0, 1, 0))
10	이브	=IF(B3="", "", IF(B3>0, 1, 0))	=IF(C3="", "", IF(C3>0, 1, 0))	=IF(D3="", "", IF(D3>0, 1, 0))	=IF(E3="", "", IF(E3>0, 1, 0))	=IF(F3="", "", IF(F3>0, 1, 0))
11	연참	=IF(B4="", "", IF(B4>0, 1, 0))	=IF(C4="", "", IF(C4>0, 1, 0))	=IF(D4="", "", IF(D4>0, 1, 0))	=IF(E4="", "", IF(E4>0, 1, 0))	=IF(F4="", "", IF(F4>0, 1, 0))
12	니모	=IF(B5="", "", IF(B5>0, 1, 0))	=IF(C5="", "", IF(C5>0, 1, 0))	=IF(D5="", "", IF(D5>0, 1, 0))	=IF(E5="", "", IF(E5>0, 1, 0))	=IF(F5="", "", IF(F5>0, 1, 0))
13	경표	=IF(B6="", "", IF(B6>0, 1, 0))	=IF(C6="", "", IF(C6>0, 1, 0))	=IF(D6="", "", IF(D6>0, 1, 0))	=IF(E6="", "", IF(E6>0, 1, 0))	=IF(F6="", "", IF(F6>0, 1, 0))

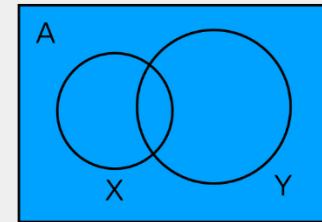
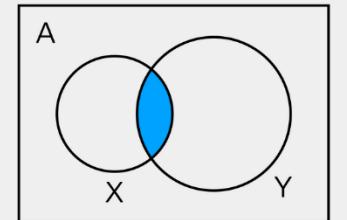
5. 기타 : 추천분석

96

- 연관분석

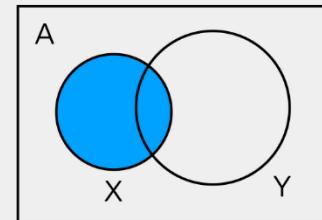
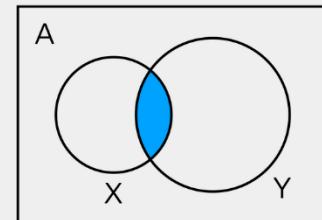
지지도 (Support)

$$\frac{|X \cap Y|}{|A|}$$



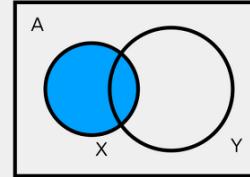
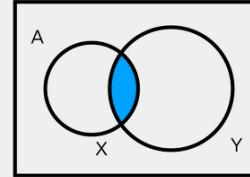
신뢰도 (Confidence)

$$\frac{|X \cap Y|}{|X|}$$



향상도 (Lift)

$$\left(\frac{\frac{|X \cap Y|}{|X|}}{\frac{|Y|}{|A|}} \right)$$



X를 구매한 사람이
Y도 구매할 확률

전체 중 Y가
판매될 확률

5. 기타 : 추천분석

97

- 연관분석

	A	B	C	D	E	F
15	I. 연관분석(장바구니분석)					
16	- 지지도(Support) : 발생 볼륨이 의미있는가?					
17	마녀2	외계인과 마녀2를 포함	2	전체수	5	0.4
18	범죄도시2	외계인과 범죄도시2를 포함	2	전체수	5	0.4
19	탑건매버릭	외계인과 탑건매버릭을 포함	4	전체수	5	0.8
20	토르러브앤썬더	외계인과 토르러브앤썬더를 포함	2	전체수	5	0.4

	A	B	C	D	E	F
15	I. 연관분석(장바구니분석)					
16	- 지지도(Support) : 발생 볼륨이 의미있는가?					
17	마녀2	= "외계인과 "&A17&"를 포함"	=SUMPRODUCT((B9:B13=1)*(D9:D13=1))	전체수	5	=C17/E17
18	범죄도시2	= "외계인과 "&A18&"를 포함"	=SUMPRODUCT((C9:C13=1)*(D9:D13=1))	전체수	5	=C18/E18
19	탑건매버릭	= "외계인과 "&A19&"를 포함"	=SUMPRODUCT((E9:E13=1)*(D9:D13=1))	전체수	5	=C19/E19
20	토르러브앤썬더	= "외계인과 "&A20&"를 포함"	=SUMPRODUCT((F9:F13=1)*(D9:D13=1))	전체수	5	=C20/E20

5. 기타 : 추천분석

- 연관분석

A	B	C	D	E	F
22	- 신뢰도(Confidence) : 외계인1를 본 사람이 함께 많이 본 것은?				
23	마녀2	외계인과 마녀2를 포함	2 외계인포함수	4	0.50
24	범죄도시2	외계인과 범죄도시2를 포함	2 외계인포함수	4	0.50
25	탑건매버릭	외계인과 탑건매버릭을 포함	4 외계인포함수	4	1.00
26	토르러브앤썬더	외계인과 토르러브앤썬더를 포함	2 외계인포함수	4	0.50

A	B	C	D	E	F
22	- 신뢰도(Confidence)				
23	마녀2	= "외계인과 "&A23&"를 포함"	=C17	외계인포함수	4 =C23/E23
24	범죄도시2	= "외계인과 "&A24&"를 포함"	=C18	외계인포함수	4 =C24/E24
25	탑건매버릭	= "외계인과 "&A25&"를 포함"	=C19	외계인포함수	4 =C25/E25
26	토르러브앤썬더	= "외계인과 "&A26&"를 포함"	=C20	외계인포함수	4 =C26/E26

5. 기타 : 추천분석

- 연관분석

A	B	C	D	E	F	G	H	I	J
28	- 향상도(Lift) : 2번째 영화의 전체 관람수 중에서 외계인1이 미친 영향은 어느 정도인가?(1초과는 연관성높음, 1은 독립적, 1미만은 연관성이적음)								
29	마녀2	외계인과 마녀2의 신뢰도	0.50	마녀2포함수	3	전체수	5	마녀2발생확률	0.6 0.83
30	범죄도시2	외계인과 범죄도시2의 신뢰도	0.50	범죄도시2포함수	4	전체수	5	범죄도시2발생확률	0.8 0.63
31	탑건매버릭	외계인과 탑건매버릭의 신뢰도	1.00	탑건매버릭포함수	4	전체수	5	탑건매버릭발생확률	0.8 1.25
32	토르러브앤썬더	외계인과 토르러브앤썬더의 신뢰도	0.50	토르러브앤썬더포함수	2	전체수	5	토르러브앤썬더발생확률	0.4 1.25

A	B	C	D	E	F	G	H	I	J
28	- 향상도(Lift) : 2번 째								
29	마녀2	= "외계인과 "&A29&"의 신뢰도"	=F23	=A29&"포함수"	3	전체수	5	=A29&"발생확률"	=E29/5 =C29/I29
30	범죄도시2	= "외계인과 "&A30&"의 신뢰도"	=F24	=A30&"포함수"	4	전체수	5	=A30&"발생확률"	=E30/5 =C30/I30
31	탑건매버릭	= "외계인과 "&A31&"의 신뢰도"	=F25	=A31&"포함수"	4	전체수	5	=A31&"발생확률"	=E31/5 =C31/I31
32	토르러브앤썬더	= "외계인과 "&A32&"의 신뢰도"	=F26	=A32&"포함수"	2	전체수	5	=A32&"발생확률"	=E32/5 =C32/I32

5. 기타 : 추천분석

100

- 코사인 유사도

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

5. 기타 : 추천분석

101

- 협업필터링 : User Based

	A	B	C	D	E	F	G	H	I	J	K
1	구분	마녀2	범죄도시2	외계인1	탑건매버릭	토르러브앤썬더					
2	Hyun		6	10	9	10					
3	이브				7	10	6				
4	연참		2	1		3	1				
5	니모			8	8	9					
6	경표		6		6	8	7				
7											
8	1. User Based										
9	User	마녀2	범죄도시2	외계인1	탑건매버릭	토르러브앤썬더	$\ A\ $	$\ B\ $	$A \cdot B$	코사인유사도	연참의외계인예상
10	Hyun		6	10	0	10	0	3.87	15.36	52	0.87
11	이브		0	0	0	10	6	3.87	11.66	36	0.80
12	연참		2	1	0	3	1	3.87	3.87	15	1.00
13	니모		0	8	0	9	0	3.87	12.04	35	0.75
14	경표		6	0	0	8	7	3.87	12.21	43	0.91

	A	B	C	D	E	F
8	1. User					
9	User	마녀2	범죄도시2	외계인1	탑건매버릭	토르러브앤썬더
10	Hyun	=IF(AND(B\$4>0,B2>0),B2,0)	=IF(AND(C\$4>0,C2>0),C2,0)	=IF(AND(D\$4>0,D2>0),D2,0)	=IF(AND(E\$4>0,E2>0),E2,0)	=IF(AND(F\$4>0,F2>0),F2,0)
11	이브	=IF(AND(B\$4>0,B3>0),B3,0)	=IF(AND(C\$4>0,C3>0),C3,0)	=IF(AND(D\$4>0,D3>0),D3,0)	=IF(AND(E\$4>0,E3>0),E3,0)	=IF(AND(F\$4>0,F3>0),F3,0)
12	연참	=IF(AND(B\$4>0,B4>0),B4,0)	=IF(AND(C\$4>0,C4>0),C4,0)	=IF(AND(D\$4>0,D4>0),D4,0)	=IF(AND(E\$4>0,E4>0),E4,0)	=IF(AND(F\$4>0,F4>0),F4,0)
13	니모	=IF(AND(B\$4>0,B5>0),B5,0)	=IF(AND(C\$4>0,C5>0),C5,0)	=IF(AND(D\$4>0,D5>0),D5,0)	=IF(AND(E\$4>0,E5>0),E5,0)	=IF(AND(F\$4>0,F5>0),F5,0)
14	경표	=IF(AND(B\$4>0,B6>0),B6,0)	=IF(AND(C\$4>0,C6>0),C6,0)	=IF(AND(D\$4>0,D6>0),D6,0)	=IF(AND(E\$4>0,E6>0),E6,0)	=IF(AND(F\$4>0,F6>0),F6,0)

	G	H	I	J	K
9	$\ A\ $	$\ B\ $	$A \cdot B$	코사인유사도	연참의외계인예상
10	=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))	=SQRT(SUM(B10^2+C10^2+D10^2+E10^2+F10^2))	=B\$12*B10+C\$12*C10+D\$12*D10+E\$12*E10+F\$12*F10	=I10/(G10*H10)	
11	=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))	=SQRT(SUM(B11^2+C11^2+D11^2+E11^2+F11^2))	=B\$12*B11+C\$12*C11+D\$12*D11+E\$12*E11+F\$12*F11	=I11/(G11*H11)	
12	=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))	=SQRT(SUM(B12^2+C12^2+D12^2+E12^2+F12^2))	=B\$12*B12+C\$12*C12+D\$12*D12+E\$12*E12+F\$12*F12	=I12/(G12*H12)	=J10*D2+J11*D3+J13*D5+J14*D6)/(J10+J11+J13+J14)
13	=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))	=SQRT(SUM(B13^2+C13^2+D13^2+E13^2+F13^2))	=B\$12*B13+C\$12*C13+D\$12*D13+E\$12*E13+F\$12*F13	=I13/(G13*H13)	
14	=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))	=SQRT(SUM(B14^2+C14^2+D14^2+E14^2+F14^2))	=B\$12*B14+C\$12*C14+D\$12*D14+E\$12*E14+F\$12*F14	=I14/(G14*H14)	

5. 기타 : 추천분석

102

- 협업필터링 : Item Based

A	B	C	D	E	F	G	H	I	J	K
구분	Hyun	이브	연참	니모	경표					
마녀2		6		2		6				
범죄도시2		10		1	8					
외계인1		9	7		8	6				
탑건매버릭		10	10	3	9	8				
토르러브앤썬더			6	1		7				
7										
8	2. Item Based									
A	B	C	D	E	F	G	H	I	J	K
User	Hyun	이브	연참	니모	경표	A	B	A*B	코사인유사도	연참의외계인예상
마녀2		6	0	0	0	6	15.17	8.49	90	0.70
범죄도시2		10	0	0	8	0	15.17	12.81	154	0.79
외계인1		9	7	0	8	6	15.17	15.17	230	1.00
탑건매버릭		10	10	0	9	8	15.17	18.57	280	0.99
토르러브앤썬더		0	6	0	0	7	15.17	9.22	84	0.60

A	B	C	D	E	F	G
2. Item Based						
User	Hyun	이브	연참	니모	경표	A
마녀2	=IF(AND(B\$4>0,B2>0),B2,0)=IF(AND(C\$4>0,C2>0),C2,0)=IF(AND(D\$4>0,D2>0),D2,0)=IF(AND(E\$4>0,E2>0),E2,0)=IF(AND(F\$4>0,F2>0),F2,0)=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))					
범죄도시2	=IF(AND(B\$4>0,B3>0),B3,0)=IF(AND(C\$4>0,C3>0),C3,0)=IF(AND(D\$4>0,D3>0),D3,0)=IF(AND(E\$4>0,E3>0),E3,0)=IF(AND(F\$4>0,F3>0),F3,0)=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))					
외계인1	=IF(AND(B\$4>0,B4>0),B4,0)=IF(AND(C\$4>0,C4>0),C4,0)=IF(AND(D\$4>0,D4>0),D4,0)=IF(AND(E\$4>0,E4>0),E4,0)=IF(AND(F\$4>0,F4>0),F4,0)=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))					
탑건매버릭	=IF(AND(B\$4>0,B5>0),B5,0)=IF(AND(C\$4>0,C5>0),C5,0)=IF(AND(D\$4>0,D5>0),D5,0)=IF(AND(E\$4>0,E5>0),E5,0)=IF(AND(F\$4>0,F5>0),F5,0)=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))					
토르러브앤썬더	=IF(AND(B\$4>0,B6>0),B6,0)=IF(AND(C\$4>0,C6>0),C6,0)=IF(AND(D\$4>0,D6>0),D6,0)=IF(AND(E\$4>0,E6>0),E6,0)=IF(AND(F\$4>0,F6>0),F6,0)=SQRT(SUM(\$B\$12^2+\$C\$12^2+\$D\$12^2+\$E\$12^2+\$F\$12^2))					

H	I	J	K
B	A*B	코사인유사도	연참의외계인예상
=SQRT(SUM(B10^2+C10^2+D10^2+E10^2+F10^2))	=B\$12*B10+C\$12*C10+D\$12*D10+E\$12*E10+F\$12*F10	=I10/(G10*H10)	
=SQRT(SUM(B11^2+C11^2+D11^2+E11^2+F11^2))	=B\$12*B11+C\$12*C11+D\$12*D11+E\$12*E11+F\$12*F11	=I11/(G11*H11)	
=SQRT(SUM(B12^2+C12^2+D12^2+E12^2+F12^2))	=B\$12*B12+C\$12*C12+D\$12*D12+E\$12*E12+F\$12*F12	=I12/(G12*H12)	=(J10*D2+J11*D3+J13*D5+J14*D6)/(J10+J11+J13+J14)
=SQRT(SUM(B13^2+C13^2+D13^2+E13^2+F13^2))	=B\$12*B13+C\$12*C13+D\$12*D13+E\$12*E13+F\$12*F13	=I13/(G13*H13)	
=SQRT(SUM(B14^2+C14^2+D14^2+E14^2+F14^2))	=B\$12*B14+C\$12*C14+D\$12*D14+E\$12*E14+F\$12*F14	=I14/(G14*H14)	

감사합니다.