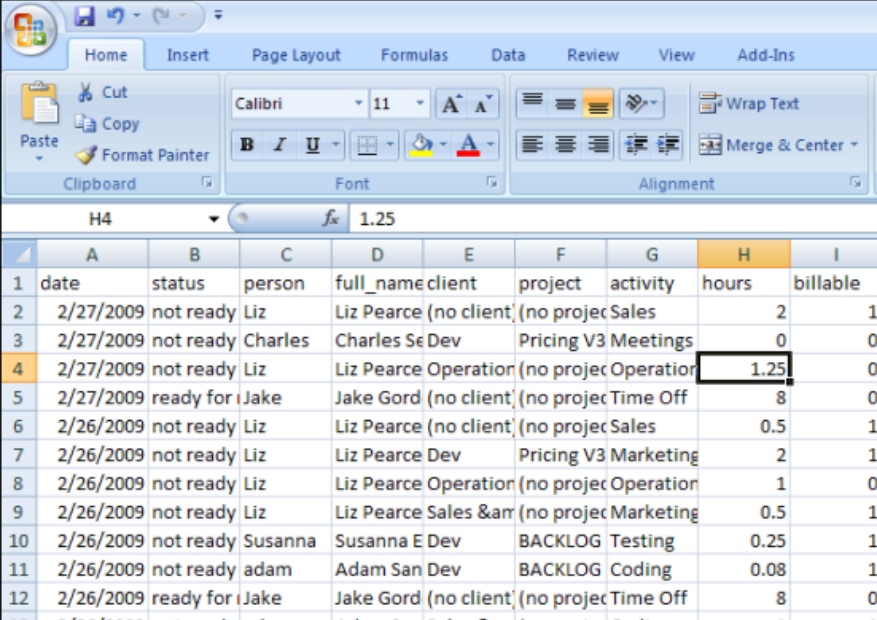


비즈니스 애널리틱스

엑셀을 활용한 기초 데이터 분석

SNU Business School

엑셀을 활용한 기초적인 데이터 분석

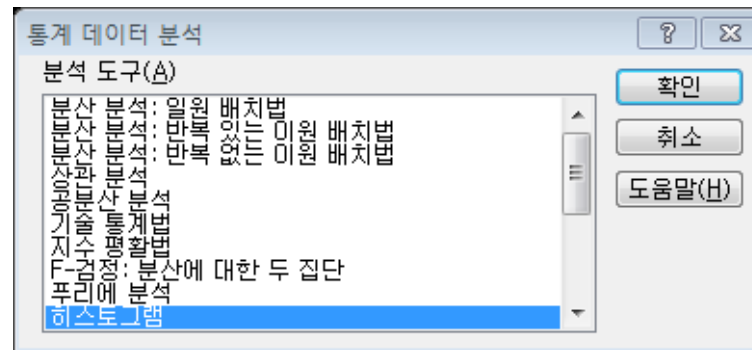


The screenshot shows the Microsoft Excel interface with the 'Home' tab selected. The ribbon includes options for Clipboard, Font, and Alignment. The active cell is H4, which contains the value 1.25. The data table below is as follows:

	A	B	C	D	E	F	G	H	I
1	date	status	person	full_name	client	project	activity	hours	billable
2	2/27/2009	not ready	Liz	Liz Pearce	(no client)	(no project)	Sales	2	1
3	2/27/2009	not ready	Charles	Charles Se	Dev	Pricing V3	Meetings	0	0
4	2/27/2009	not ready	Liz	Liz Pearce	Operation	(no project)	Operation	1.25	0
5	2/27/2009	ready for	Jake	Jake Gord	(no client)	(no project)	Time Off	8	0
6	2/26/2009	not ready	Liz	Liz Pearce	(no client)	(no project)	Sales	0.5	1
7	2/26/2009	not ready	Liz	Liz Pearce	Dev	Pricing V3	Marketing	2	1
8	2/26/2009	not ready	Liz	Liz Pearce	Operation	(no project)	Operation	1	0
9	2/26/2009	not ready	Liz	Liz Pearce	Sales & an	(no project)	Marketing	0.5	1
10	2/26/2009	not ready	Susanna	Susanna E	Dev	BACKLOG	Testing	0.25	1
11	2/26/2009	not ready	adam	Adam San	Dev	BACKLOG	Coding	0.08	1
12	2/26/2009	ready for	Jake	Jake Gord	(no client)	(no project)	Time Off	8	0

엑셀 분석 도구

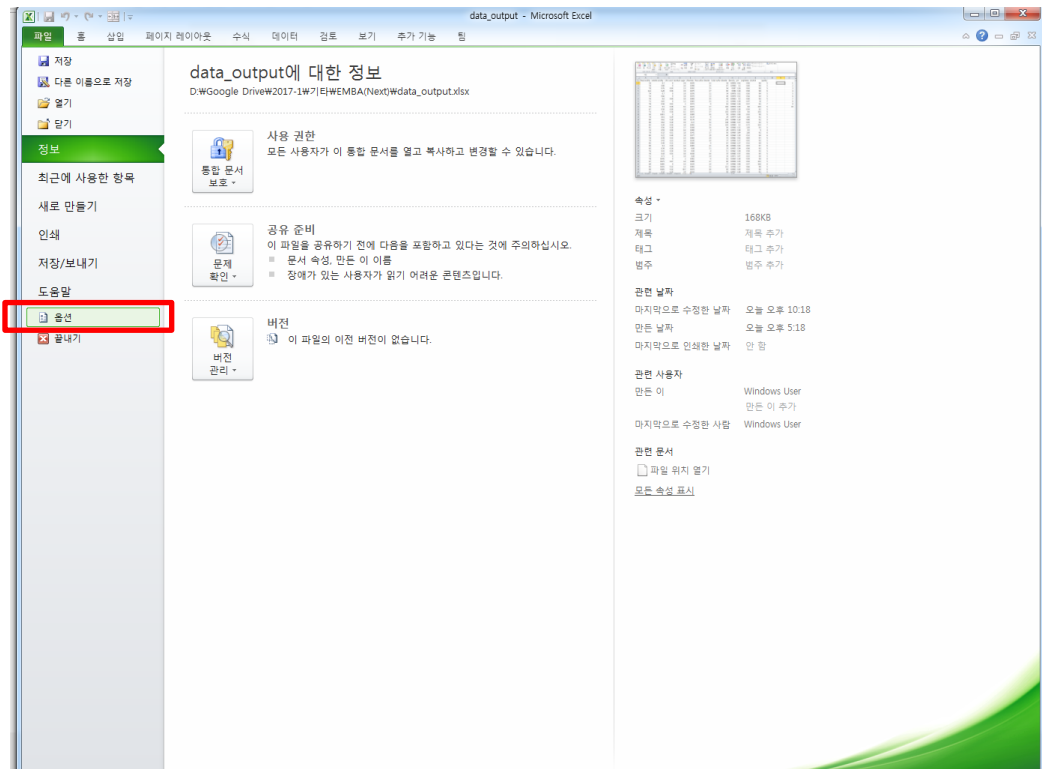
- 엑셀의 추가 기능 중 ‘분석 도구’를 활용하면 기초적인 데이터 분석을 엑셀을 활용하여 손쉽게 할 수 있다
 - 분산 분석, 상관 분석, 기술 통계량 계산, F-검정, 푸리에 분석 등 다양한 분석을 클릭 몇 번으로 간단히 수행할 수 있음
- 엑셀의 추가 기능에 기본적으로 포함되어 있으므로 따로 다운받을 필요 없이 활성화만 해주면 된다



엑셀 분석 도구

· 엑셀 분석 도구 활성화하기

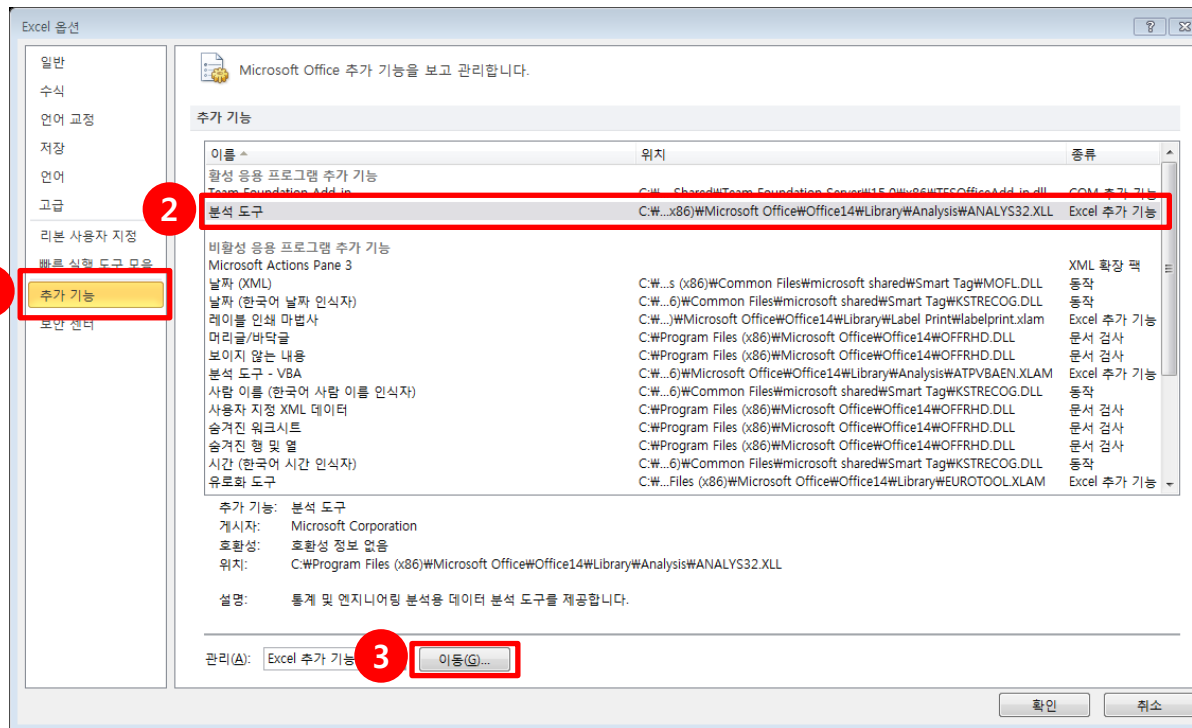
- [파일] 탭에서 [옵션]을 클릭한다



엑셀 분석 도구

· 엑셀 분석 도구 활성화하기

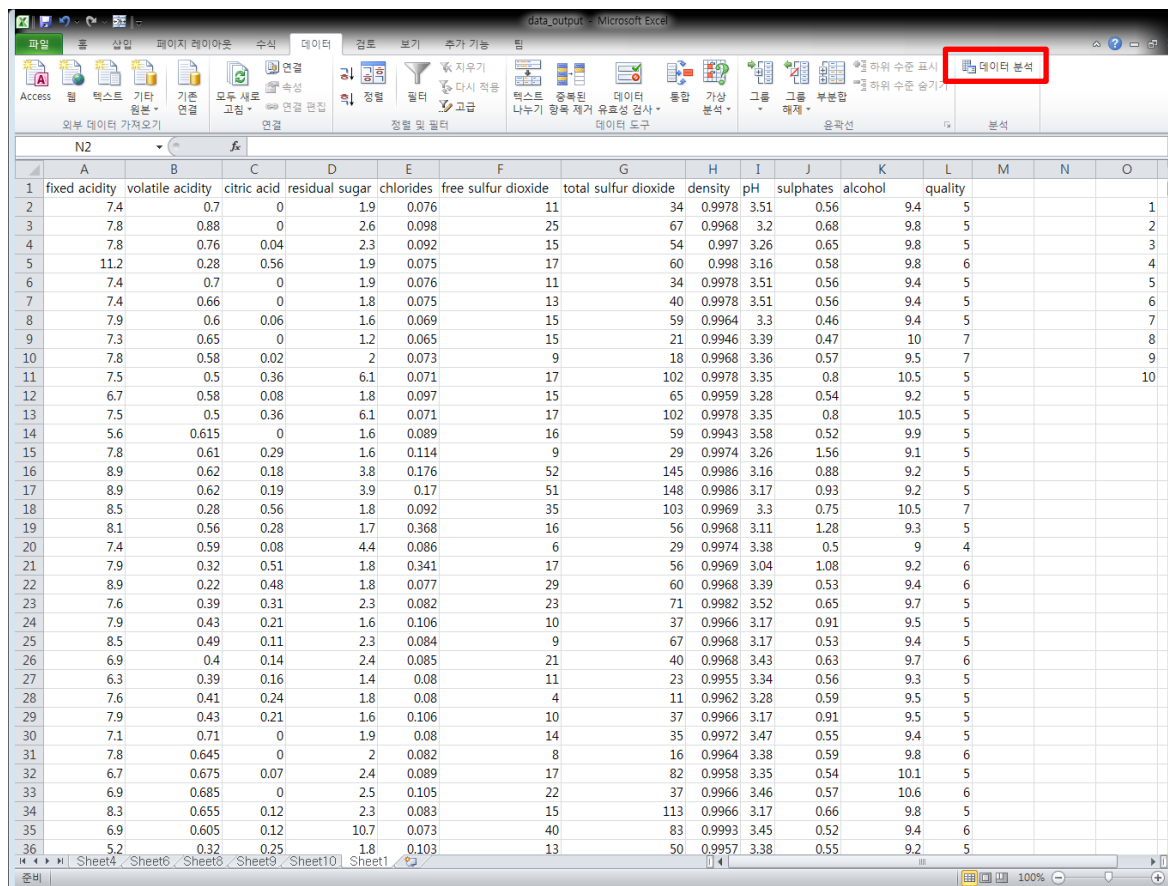
- [추가 기능] 탭에서 분석 도구를 선택하고 아래에서 [이동(G)]를 클릭한다



엑셀 분석 도구

· 엑셀 분석 도구 활성화하기

- [데이터] 탭에서 [데이터 분석]이 추가된 것을 볼 수 있다



기술통계량 분석하기

· 기술(記述) 통계: 측정이나 실험에서 수집한 자료의 정리, 표현, 요약, 해석 등을 통해 자료의 특성을 규명하는 통계적 방법[source: <https://ko.wikipedia.org/wiki/%ED%86%B5%EA%B3%84%ED%95%99>]

- 표본의 특성을 나타내는 대표값인 평균, 분산, 중앙값 등을 나타내는 기본적인 기술 통계량 분석 뿐만 아니라, 모집단에서 어떤 인자들이 있는지 뽑아보는 인자분석, 특정표본이 어떤 모집단에 속하는지 판단하는 판별분석 등이 있다

- 평균, 중앙값, 그리고 최빈값

■ 평균: 기대값이라고도 하며, 모든 표본값을 더한 후 표본값의 개수(n)으로 나눈 값. 가장 흔하게 활용되며 계산하기 쉽다는 장점이 있지만, 너무 크거나 작은 이상치(outlier)에 취약하다는 단점이 있다. 일반적으로, 표본의 분산(편차)이 지나치게 큰 경우, 평균은 좋은 대표값이 되지 않는다

■ 중앙값: 큰 값부터 작은 값까지 일렬로 늘어놓았을 때 중간에 있는 값

■ 최빈값: 표본값 중 가장 많이 나오는 값. 그렇지만 현실에서 표본들이 정확히 같은 값을 가지는 경우가 많지 않아 활용하기 쉽지 않다

■ 결국 데이터의 형태를 보고 가장 적절한 대표값을 문맥에 맞게 찾아가는 것이 중요!!

기술통계량 분석하기

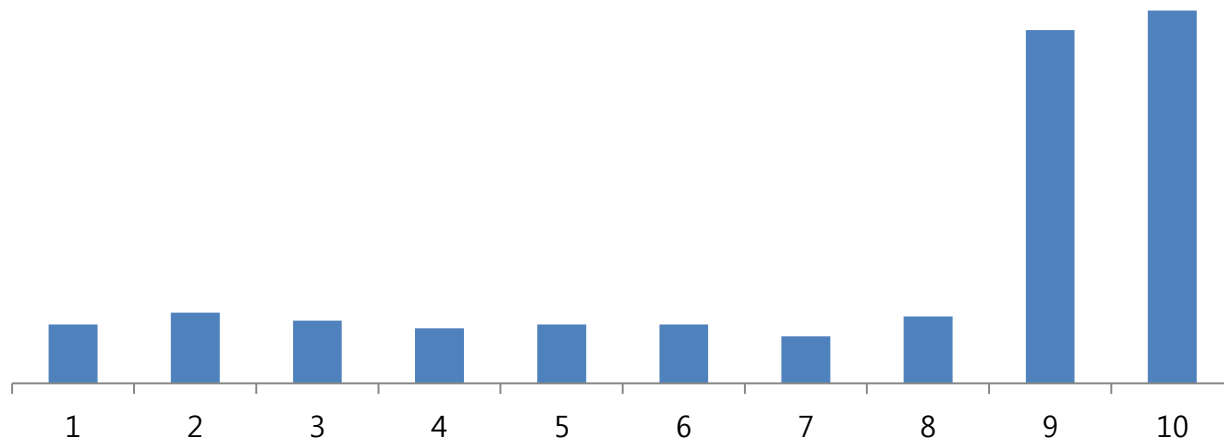
· 기술(記述) 통계: 측정이나 실험에서 수집한 자료의 정리, 표현, 요약, 해석 등을 통해 자료의 특성을 규명하는 통계적 방법[source: <https://ko.wikipedia.org/wiki/%ED%86%B5%EA%B3%84%ED%95%99>]

- 예시: 평균, 중앙값, 그리고 최빈값[source: <https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php>]

■ 예를 들어, 어떤 부서의 직원들의 연봉 정보가 아래와 같다고 하자(단위: 천만 원)

직원 ID	1	2	3	4	5	6	7	8	9	10
연봉	1.5	1.8	1.6	1.4	1.5	1.5	1.2	1.7	9	9.5

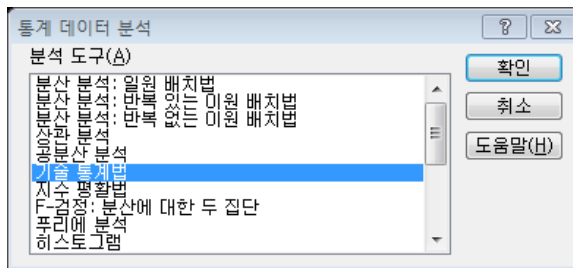
■ 이 부서의 직원들의 연봉을 평균을 내보면 3천70만원이라는 값이 나온다. 하지만 데이터를 조금만 자세히 보면 알 수 있듯이, 3000만원 근방의 값을 받는 직원은 한 명도 없다. 이 경우, 직원들의 연봉의 대표값을 산정하기 위해 평균을 활용하는 것은 적절해 보이지 않는다



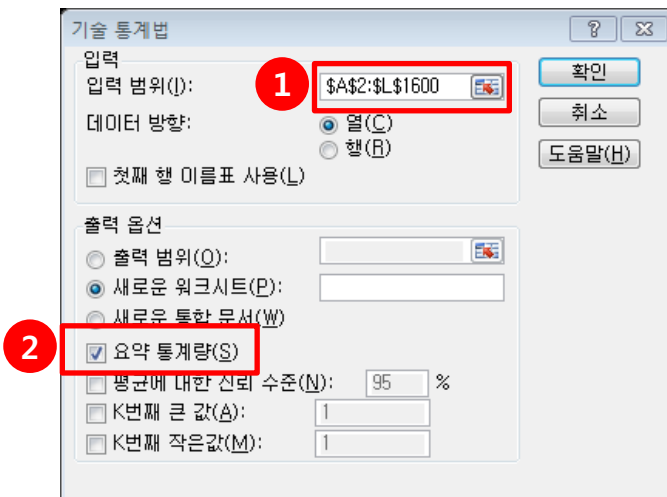
기술통계량 분석하기

- 데이터의 기술통계량(평균, 표준오차, 중앙값, 최빈값 등)을 계산할 수 있다

- [데이터 분석] 에서 [기술 통계법]을 선택한 후 [확인]을 누른다



- 기술 통계량을 계산하고 싶은 범위를 선택한 뒤 [요약 통계량(S)]를 체크하고 [확인]을 누른다



기술통계량 분석하기

· 아래와 같이 각 열(column)마다 기술통계량이 계산되어 나온다

- 참고: 기술통계량의 입력 범위에는 숫자만 들어가야 하므로 각 열의 이름인 헤더(header)가 범위에 들어가지 않게 유의한다

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11	Column12														
3	평균	8.319637	평균	0.527821	평균	0.270976	평균	2.538806	평균	0.087467	평균	15.87492	평균	46.46779	평균	0.996747	평균	3.311113	평균	0.658149	평균	10.42298	평균	5.636023		
4	표준 오차	0.043541	표준 오차	0.004478	표준 오차	0.004872	표준 오차	0.035259	표준 오차	0.001177	표준 오차	0.261586	표준 오차	0.82264	표준 오차	4.72E-05	표준 오차	0.003861	표준 오차	0.004239	표준 오차	0.02665	표준 오차	0.020196		
5	중앙값	7.9	중앙값	0.52	중앙값	0.26	중앙값	2.2	중앙값	0.079	중앙값	14	중앙값	38	중앙값	0.99675	중앙값	3.31	중앙값	0.62	중앙값	10.2	중앙값	6		
6	최빈값	7.2	최빈값	0.6	최빈값	0	최빈값	2	최빈값	0.08	최빈값	6	최빈값	28	최빈값	0.9972	최빈값	3.3	최빈값	0.6	최빈값	9.5	최빈값	5		
7	표준 편차	1.741096	표준 편차	0.17906	표준 편차	0.194801	표준 편차	1.409928	표준 편차	0.047065	표준 편차	10.46016	표준 편차	32.89532	표준 편차	0.001887	표준 편차	0.154386	표준 편차	0.169507	표준 편차	1.065668	표준 편차	0.807569		
8	분산	3.031416	분산	0.032062	분산	0.037947	분산	1.987897	분산	0.002215	분산	109.4149	분산	1082.102	분산	3.56E-06	분산	0.023835	분산	0.028733	분산	1.135647	분산	0.652168		
9	첨도	1.132143	첨도	1.225542	첨도	-0.789	첨도	28.6176	첨도	41.71579	첨도	2.023562	첨도	3.809824	첨도	0.934079	첨도	0.806943	첨도	11.72025	첨도	0.200029	첨도	0.296708		
10	왜도	0.982751	왜도	0.671593	왜도	0.318337	왜도	4.540655	왜도	5.680347	왜도	1.250567	왜도	1.515531	왜도	0.071288	왜도	0.193683	왜도	2.428672	왜도	0.860829	왜도	0.217802		
11	범위	11.3	범위	1.46	범위	1	범위	14.6	범위	0.599	범위	71	범위	283	범위	0.01362	범위	1.27	범위	1.67	범위	6.5	범위	5		
12	최소값	4.6	최소값	0.12	최소값	0	최소값	0.9	최소값	0.012	최소값	1	최소값	6	최소값	0.99007	최소값	2.74	최소값	0.33	최소값	8.4	최소값	3		
13	최대값	15.9	최대값	1.58	최대값	1	최대값	15.5	최대값	0.611	최대값	72	최대값	289	최대값	1.00369	최대값	4.01	최대값	2	최대값	14.9	최대값	8		
14	합	13303.1	합	843.985	합	433.29	합	4059.55	합	139.859	합	25384	합	74302	합	1593.798	합	5294.47	합	1052.38	합	16666.35	합	9012		
15	관측수	1599	관측수	1599	관측수	1599	관측수	1599	관측수	1599	관측수	1599	관측수	1599	관측수	1599	관측수	1599	관측수	1599	관측수	1599	관측수	1599		

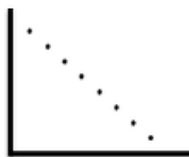
상관 분석

· 상관계수(correlation coefficient): 두 변수 간에 어떤 선형적 관계(linear relationship)를 가지고 있는지를 나타내는 방법[source: <https://ko.wikipedia.org/wiki/%EC%83%81%EA%B4%80%EB%B6%84%EC%84%9D>]

- 일반적으로 피어슨 상관계수(Pearson correlation coefficient)가 주로 활용되며, 최소값은 -1 (완벽한 음의 상관관계), 최대값은 1 (완벽한 양의 상관관계)이다.

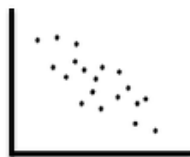
$$r(X, Y) = \frac{X \text{와 } Y \text{가 함께 변하는 정도}}{X \text{와 } Y \text{가 따로 변하는 정도}}$$

■ 한 가지 주의할 점은, 상관계수는 선형 관계만 나타내므로, 비선형(nonlinear) 관계는 파악하기 힘들다. 상관계수가 0이더라도 변수 간의 비선형 관계는 존재할 수 있다



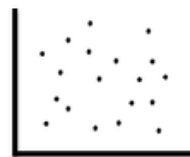
$r = -1$

음의 상관관계가
강하다.



$-1 < r < 0$

음의 상관관계가
있기는 하다.



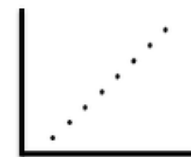
$r = 0$

상관관계가 없다.



$0 < r < 1$

양의 상관관계가
있기는 하다.

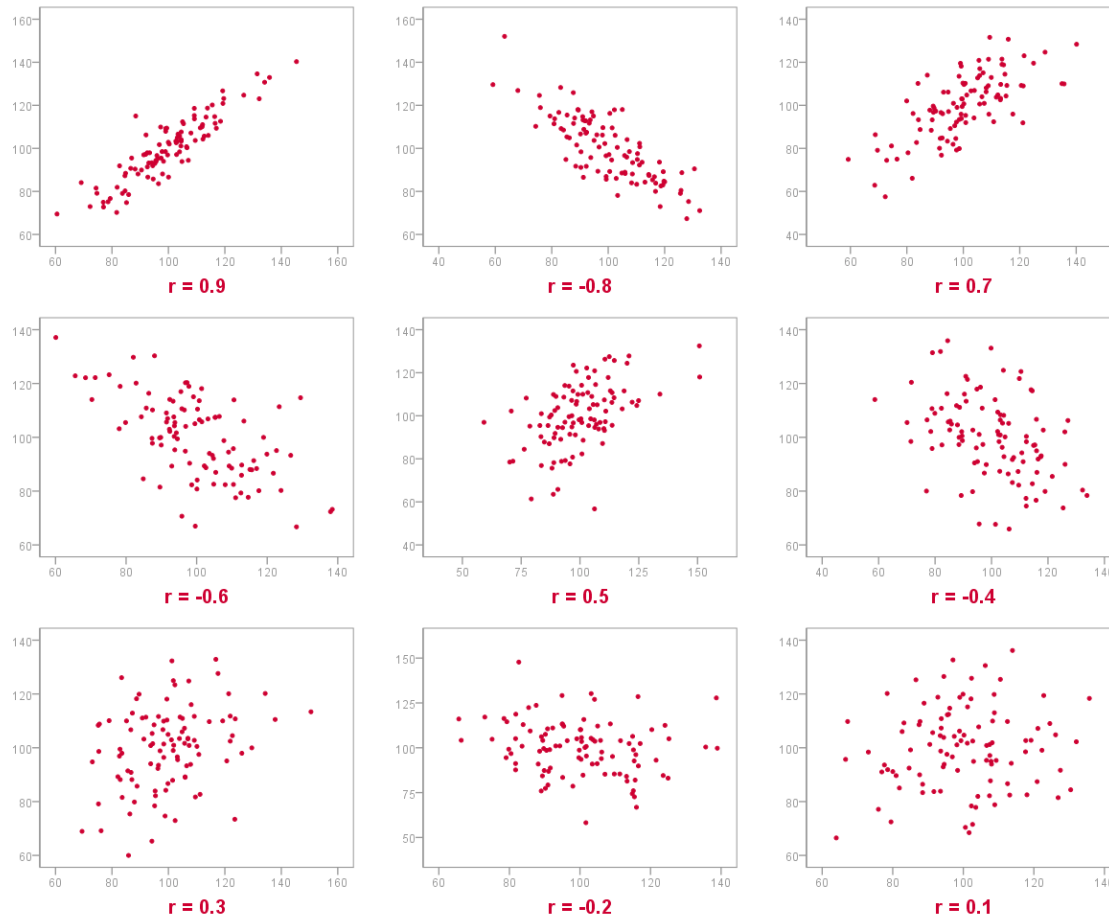


$r = +1$

양의 상관관계가
강하다.

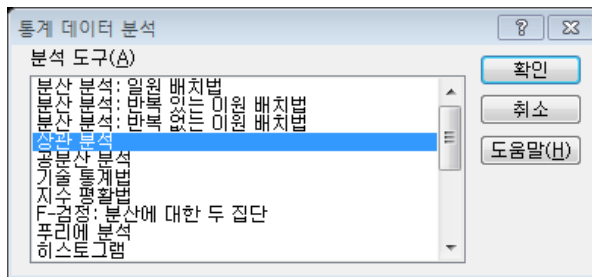
상관 분석

- 상관계수(correlation coefficient): 두 변수 간에 어떤 선형적 관계(linear relationship)를 가지고 있는지를 나타내는 방법[source: <https://ko.wikipedia.org/wiki/%EC%83%81%EA%B4%80%EB%B6%84%EC%84%9D>]

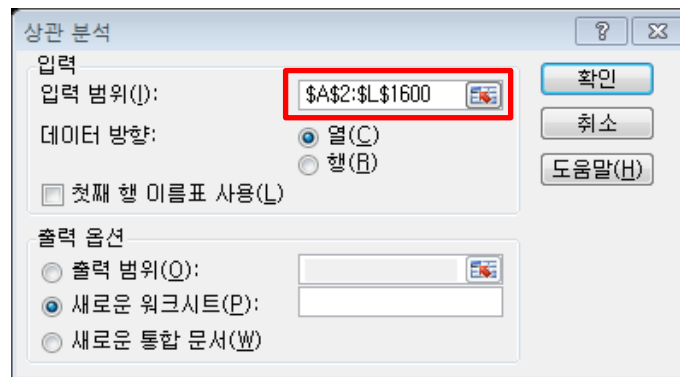


상관 분석

- 데이터의 상관 관계(correlation)을 계산할 수 있다
- [데이터 분석] 에서 [상관 분석]을 선택한 후 [확인]을 누른다

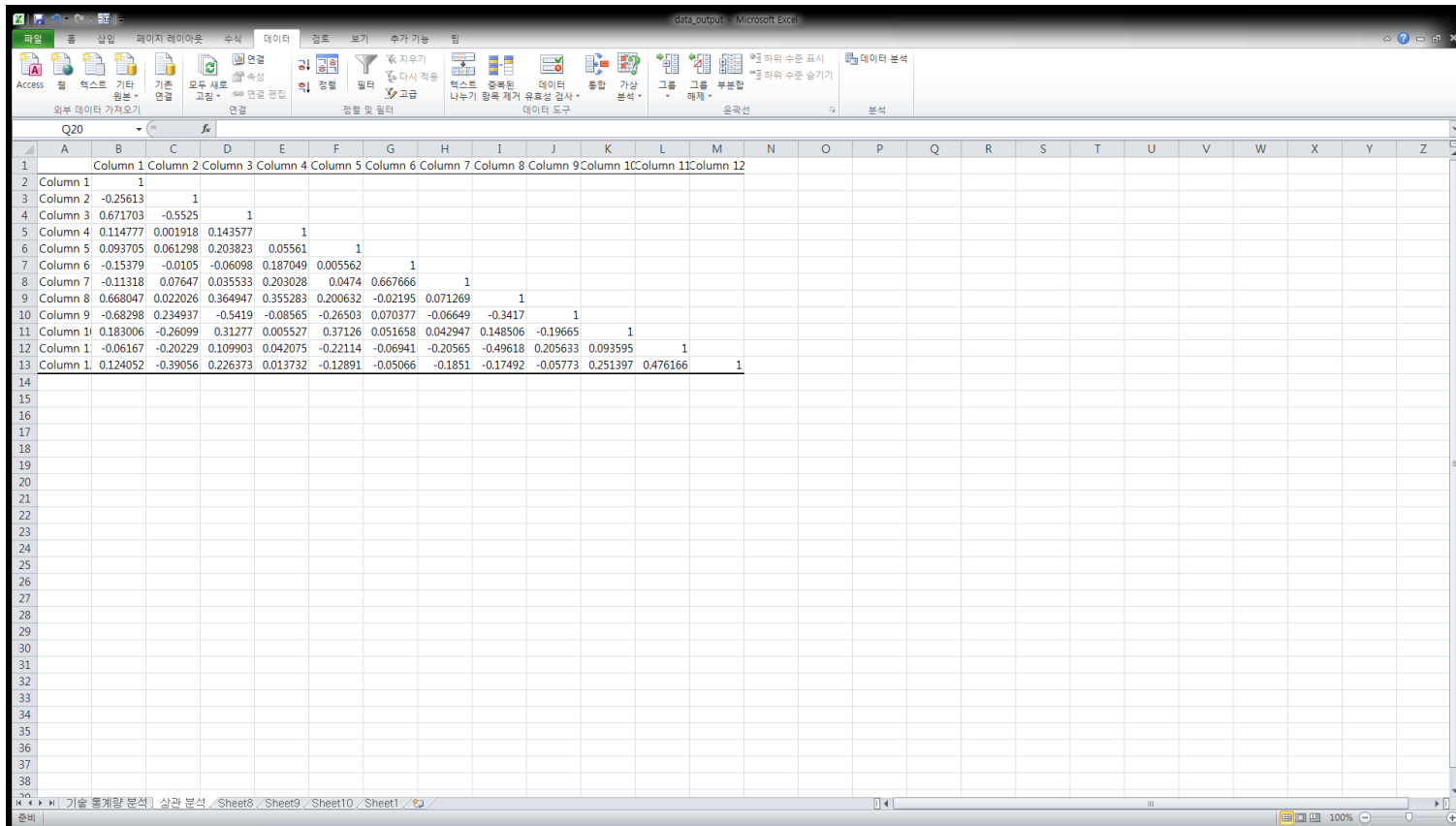


- 상관 분석을 수행하고 싶은 범위를 선택한 뒤 [확인]을 누른다



상관 분석

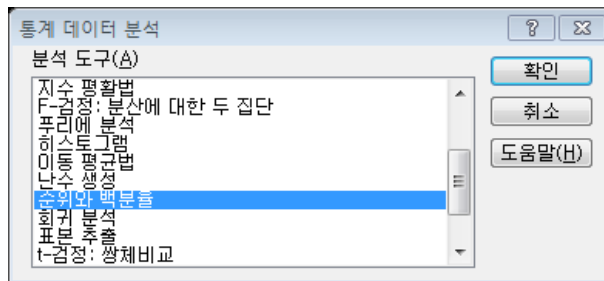
- 아래와 같이 각 열 별로 상관계수(correlation coefficient)가 출력된 것을 볼 수 있다
- 어떤 변수들 끼리 높고 낮은 상관관계를 갖는지 한 눈에 알아볼 수 있다



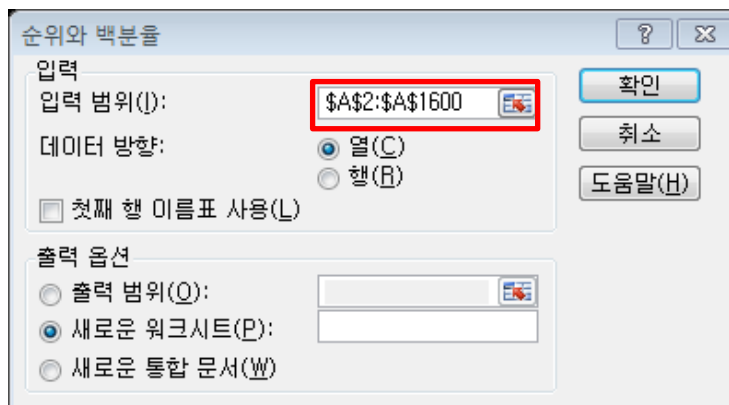
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12
Column 1	1											
Column 2	-0.25613	1										
Column 3	0.671703	-0.5525	1									
Column 4	0.114777	0.001918	0.143577	1								
Column 5	0.093705	0.061298	0.203823	0.05561	1							
Column 6	-0.15379	-0.0105	-0.06098	0.187049	0.005562	1						
Column 7	-0.11318	0.07647	0.035533	0.203028	0.0474	0.667666	1					
Column 8	0.668047	0.022026	0.364947	0.355283	0.200632	-0.02195	0.071269	1				
Column 9	-0.68298	0.234937	-0.5419	-0.08565	-0.26503	0.070377	-0.06649	-0.3417	1			
Column 10	0.183006	-0.26099	0.31277	0.005527	0.37126	0.051658	0.042947	0.148506	-0.19665	1		
Column 11	-0.06167	-0.20229	0.109903	0.042075	-0.22114	-0.06941	-0.20565	-0.49618	0.205633	0.093595	1	
Column 12	0.124052	-0.39056	0.226373	0.013732	-0.12891	-0.05066	-0.1851	-0.17492	-0.05773	0.251397	0.476166	1

순위와 백분율

- 각 데이터 포인트의 순위와 백분율을 계산할 수 있다
- [데이터 분석] 에서 [순위와 백분율]을 선택한 후 [확인]을 누른다



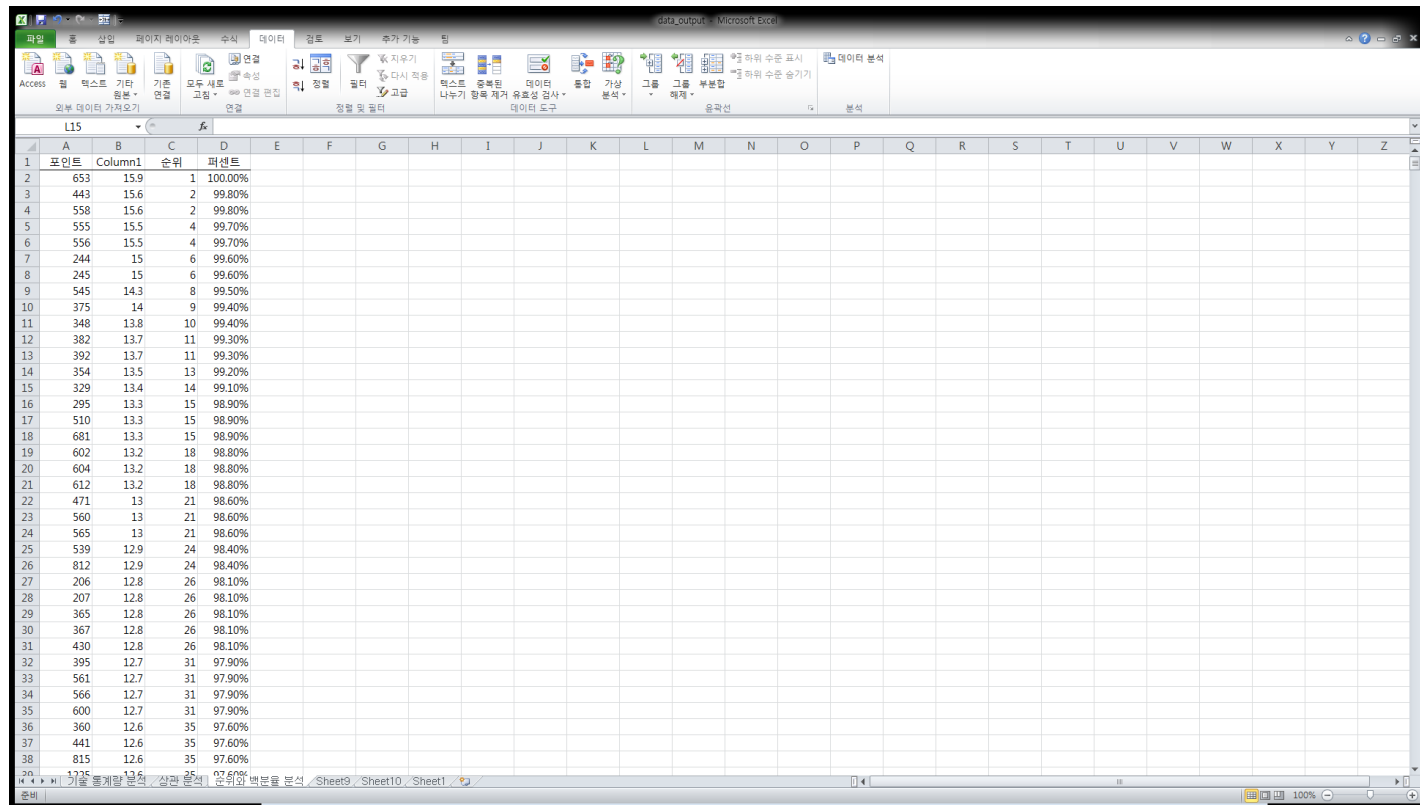
- 순위와 백분율 분석을 수행하고 싶은 범위를 선택한 뒤 [확인]을 누른다



순위와 백분율

· 아래와 같이 순위 및 백분율이 출력되는 것을 볼 수 있다

- 참고: 75%가 마크된 포인트가 일반적으로 말하는 3/4 분위수(third quartile), 50%가 마크된 포인트가 2/4 분위수 혹은 중앙값(median), 25%가 마크된 포인트가 1/4 분위수(first quartile)임



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	포인트	Column1	순위	퍼센트																						
1	653	15.9	1	100.00%																						
2	443	15.6	2	99.80%																						
3	558	15.6	2	99.80%																						
4	555	15.5	4	99.70%																						
5	556	15.5	4	99.70%																						
6	244	15	6	99.60%																						
7	245	15	6	99.60%																						
8	545	14.3	8	99.50%																						
9	375	14	9	99.40%																						
10	348	13.8	10	99.40%																						
11	382	13.7	11	99.30%																						
12	392	13.7	11	99.30%																						
13	354	13.5	13	99.20%																						
14	329	13.4	14	99.10%																						
15	295	13.3	15	98.90%																						
16	510	13.3	15	98.90%																						
17	681	13.3	15	98.90%																						
18	602	13.2	18	98.80%																						
19	604	13.2	18	98.80%																						
20	612	13.2	18	98.80%																						
21	471	13	21	98.60%																						
22	560	13	21	98.60%																						
23	565	13	21	98.60%																						
24	539	12.9	24	98.40%																						
25	812	12.9	24	98.40%																						
26	206	12.8	26	98.10%																						
27	207	12.8	26	98.10%																						
28	365	12.8	26	98.10%																						
29	367	12.8	26	98.10%																						
30	430	12.8	26	98.10%																						
31	395	12.7	31	97.90%																						
32	561	12.7	31	97.90%																						
33	566	12.7	31	97.90%																						
34	600	12.7	31	97.90%																						
35	360	12.6	35	97.60%																						
36	441	12.6	35	97.60%																						
37	815	12.6	35	97.60%																						
38	1235	13.3	25	97.60%																						

도수 분포표와 히스토그램

· 도수 분포표와 히스토그램을 손쉽게 그릴 수 있다

- 우선, 도수분포표의 각 계급을 적는다(여기서는 와인 품질의 도수분포표와 히스토그램을 그리기 위해 1부터 10까지의 정수로 계급을 표현함)

N	O	P
	1	
	2	
	3	
	4	
1	5	
	6	
	7	
	8	
	9	
	10	

도수 분포표와 히스토그램

· 도수분포표와 히스토그램

- 도수분포표: 측정값을 몇 개의 계급으로 나누고, 각 계급에 속한 도수를 조사하여 나타낸 것

■ 계급 간의 상대적인 도수를 비교하기 쉽다

학생들의 키

(단위 : cm)

144	168	148	129
162	130	153	154
167	135	128	140
134	159	149	145
138	151	146	150

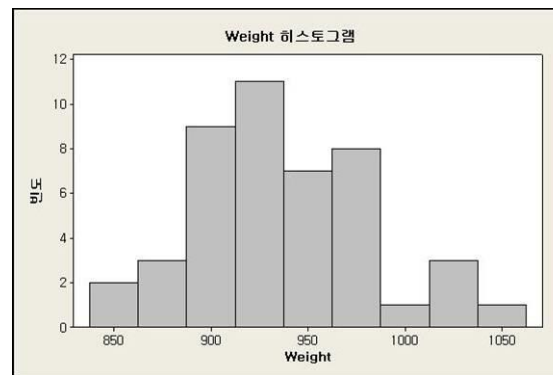
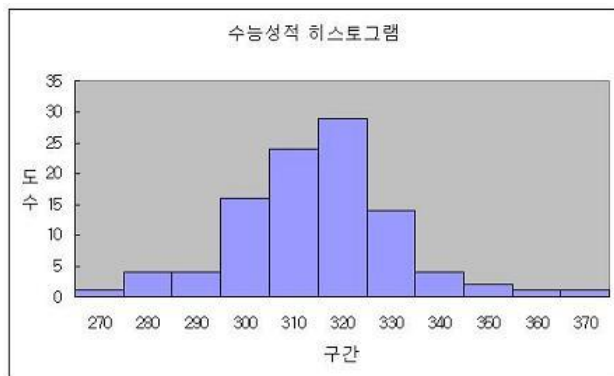
<정리되지 않은 자료>

학생들의 키

키 (cm)	학생 수(명)
120 이상 ~ 130 미만	2
130 ~ 140	4
140 ~ 150	6
150 ~ 160	5
160 ~ 170	3
합 계	20

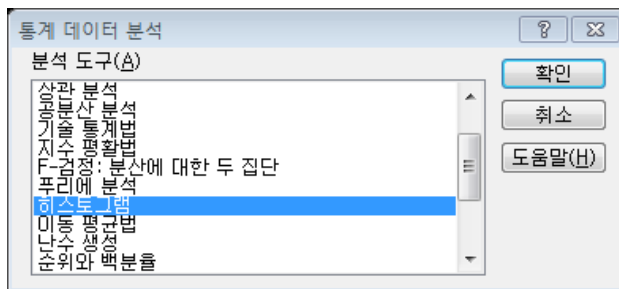
<도수분포표>

- 히스토그램: 가로축에 계급을, 세로축에 도수를 취하고, 도수분포표의 상태를 직사각형의 기둥 모양으로 나타낸 그래프

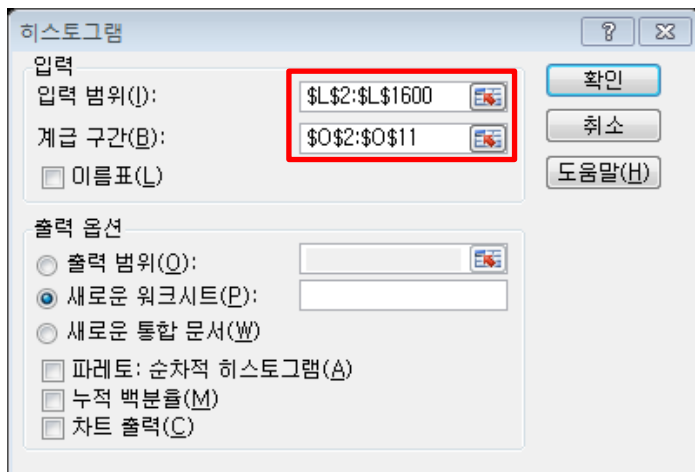


도수 분포표와 히스토그램

- 도수 분포표와 히스토그램을 손쉽게 그릴 수 있다
- [데이터 분석] 에서 [히스토그램]을 선택한 후 [확인]을 누른다

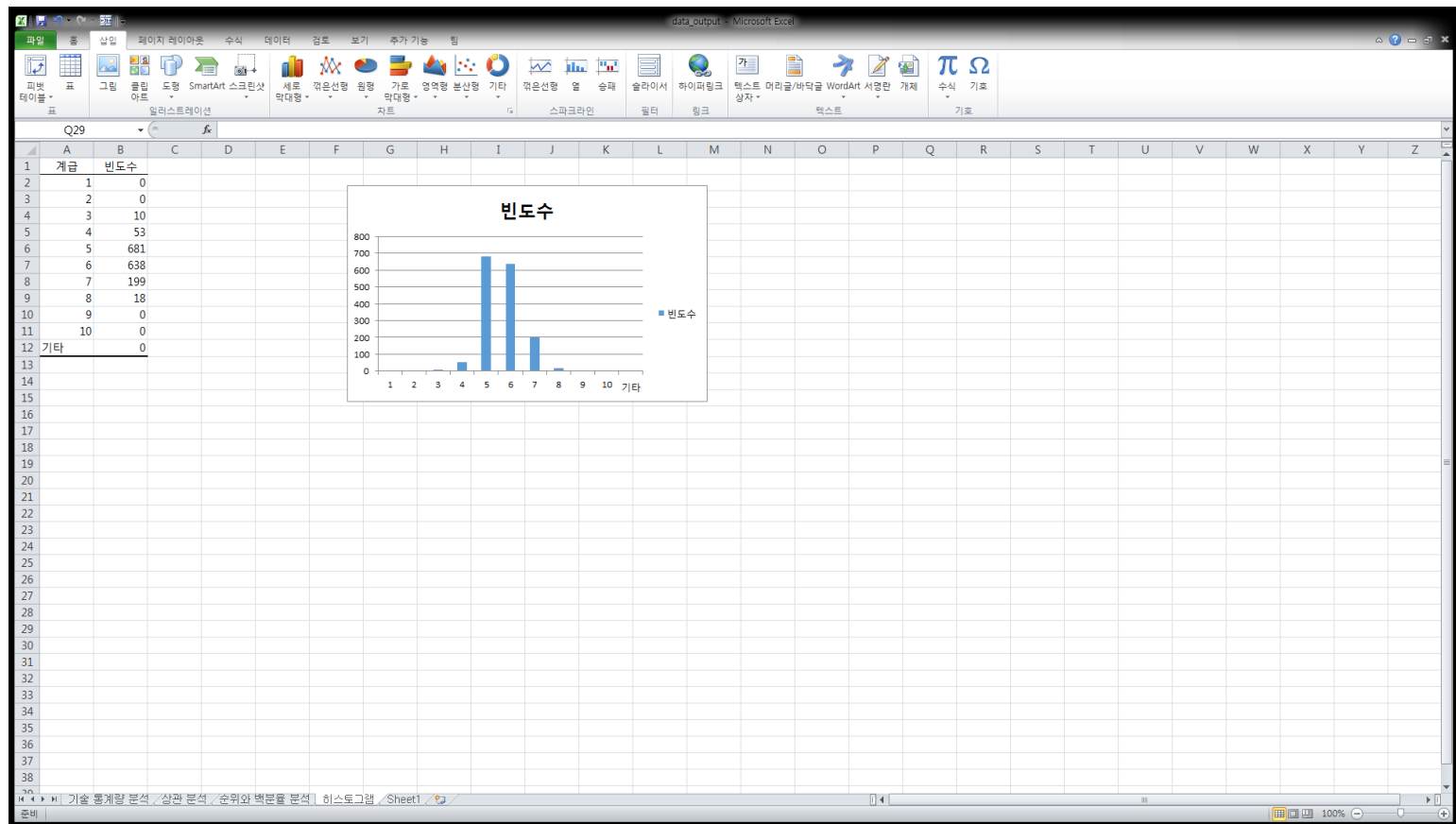


- 입력 범위와 계급 구간이 있는 곳을 선택하고 [확인]을 누른다



도수 분포표와 히스토그램

- 아래와 같이 도수 분포표가 그려지고 이를 바탕으로 히스토그램을 그릴 수 있다
- 세로 막대형 차트를 도수 분포표를 가지고 그리면 히스토그램이 된다



실습

- Kaggle 혹은 UCI 데이터 중 수치형 데이터를 가져와 엑셀을 활용해 간단한 분석을 해본다

The screenshot shows a Microsoft Excel spreadsheet titled 'car_sales.xlsx'. The data is organized in columns: A (manufact), B (model), C (sales), D (resale), E (price), F (type), G (engine_s), H (horsepow), I (wheelbas), J (width), K (length), L (curb_wgt), M (fuel_cap), and N (mpg). The rows list various car models and their corresponding specifications. A 'Data Analysis' dialog box is open, displaying a list of analysis tools including Anova: Single Factor, Anova: Two-Factor With Replication, Anova: Two-Factor Without Replication, Correlation, Covariance, Descriptive Statistics, Exponential Smoothing, F-Test Two-Sample for Variances, Fourier Analysis, and Histogram. The 'Correlation' option is highlighted.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	manufact	model	sales	resale	price	type	engine_s	horsepow	wheelbas	width	length	curb_wgt	fuel_cap	mpg
2	Acura	Integra	16.919	16.36	21.5	0	1.8	140	101.2	67.3	172.4	2.639	13.2	28
3	Acura	TL	39.384	19.875	28.4	0	3.2	225	108.1	70.3	192.9	3.517		
4	Acura	RL	8.588	29.725	42	0	3.5	210	114.6	71.4	196.6	3.85		
5	Audi	A4	20.397	22.255	23.99	0	1.8	150	102.6	68.2	178	2.998		
6	Audi	A6	18.78	23.555	33.95	0	2.8	200	108.7	76.1	192	3.561		
7	Audi	A8	1.38	39	62	0	4.2	310	113	74	198.2	3.902		
8	BMW	323i	19.747		26.99	0	2.5	170	107.3	68.4	176	3.179		
9	BMW	328i	9.231	28.675	33.4	0	2.8	193	107.3	68.5	176	3.197		
10	BMW	528i	17.527	36.125	38.9	0	2.8	193	111.4	70.9	188	3.472		
11	Buick	Century	91.561	12.475	21.975	0	3.1	175	109	72.7	194.6	3.368		
12	Buick	Regal	39.35	13.74	25.3	0	3.8	240	109	72.7	196.2	3.543		
13	Buick	Park Avenue	27.851	20.19	31.965	0	3.8	205	113.8	74.7	206.8	3.778		
14	Buick	LeSabre	83.257	13.36	27.885	0	3.8	205	112.2	73.5	200	3.591	17.5	25
15	Cadillac	DeVille	63.729	22.525	39.895	0	4.6	275	115.3	74.5	207.2	3.978	18.5	22
16	Cadillac	Eldorado	6.536	25.725	39.665	0	4.6	275	108	75.5	200.6	3.843	19	22
17	Cadillac	Catera	11.185	18.225	31.01	0	3	200	107.4	70.3	194.8	3.77	18	22
18	Chevrolet	Cavalier	145.519	9.25	13.26	0	2.2	115	104.1	67.9	180.9	2.676	14.3	27
19	Chevrolet	Malibu	135.126	11.225	16.535	0	3.1	170	107	69.4	190.4	3.051	15	25
20	Chevrolet	Lumina	24.629	10.31	18.89	0	3.1	175	107.5	72.5	200.9	3.33	16.6	25
21	Chevrolet	Monte Carlo	42.593	11.525	19.39	0	3.4	180	110.5	72.7	197.9	3.34	17	27
22	Chevrolet	Camaro	26.402	13.025	24.34	0	3.8	200	101.1	74.1	193.2	3.5	16.8	25
23	Chevrolet	Corvette	17.947	36.225	45.705	0	5.7	345	104.5	73.6	179.7	3.21	19.1	22
24	Chevrolet	Prizm	32.299	9.125	13.96	0	1.8	120	97.1	66.7	174.3	2.398	13.2	33

References

- <http://www.excel-easy.com/>
- <https://support.office.com/ko-kr/>